Université de Montréal

# Phylogénie et transferts horizontaux de gènes chez les bactéries

par
Fabrice Baro

Faculté des études supérieures
Programme de bio-informatique

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en bio-informatique

Avril, 2007

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

**Phylogénie et transferts horizontaux de gènes chez les bactéries**

présenté par:

Fabrice Baro

a été évalué par un jury composé des personnes suivantes:

| | |
|---|---|
| B. Franz Lang, | président-rapporteur |
| Hervé Philippe, | directeur de recherche |
| Sylvie Hamel, | membre du jury |

**Mémoire accepté le:** .....

## RÉSUMÉ

La phylogénie a pour but la recherche des relations de parenté entre les organismes. Plusieurs phénomènes perturbent cette recherche, parmi lesquels les transferts horizontaux de gènes, qui consistent essentiellement à l'échange de matériel génétique entre des individus d'espèces différentes. Récemment, de nombreux cas de transferts horizontaux ont été mis en évidence, même entre des espèces éloignées. Cela nous amène à nous demander s'il est pertinent de parler d'un arbre du vivant : est-ce que tous les gènes sont potentiellement transférables horizontalement, ou bien exite-t-il des gènes qui sont peu ou pas transférés, permettant ainsi l'obtention de la vraie phylogénie des organismes ? Notre hypothèse est que les gènes qui sont les plus répandus seront moins transférés que ceux qui sont peu répandus. Pour tester notre hypothèse, nous recherchons les gènes homologues non paralogues (HNP) dans les génomes d'un certain nombre de bactéries en faisant une recherche des hits BLAST réciproques. On récolte les gènes HNP présents dans tous nos génomes, tous les génomes sauf un, etc. Pour chacun de ces groupes de gènes, on reconstruit l'arbre phylogénétique puis on le compare avec l'arbre original des organismes. En comparant les branches de ces arbres, nous seront capables d'observer s'il y a eu des tranferts entre les espèces étudiées.



**Mots clés : phylogénie, transferts horizontaux de gènes, procaryotes, évolution moléculaire, génomique, phylogénomique.**

## ABSTRACT

Phylogeny's goal is to find the relationships between organisms. Many obstacles can deter this search, one of them being horizontal gene transfer (HGT). HGT is the exchange of genetic material between different species. Recently, many cases of HGT have been discovered, even between distantly related species. This brings us to question the concept of the tree of life: are any genes susceptible to HGT, or are there genes that are seldom or not transferred, allowing the true phylogeny to be found ? Our hypothesis is that widespread genes are less prone to be transferred than rare ones. We identify non-paralog homologous (HNP) genes in a limited number of completely sequenced bacterial genomes. We select those who are present in all genomes, all but one, etc. We infer the phylogeny for each HNP gene family, then we compare it to the organismal phylogeny in order to determine whether some HGT has taken place.



**Keywords: phylogeny, horizontal gene transfer, prokaryotes, molecular evolution, genomics, phylogenomics.**

## TABLE DES MATIÈRES


**RÉSUMÉ** . . . . . **iii**

**ABSTRACT** . . . . . **iv**

**TABLE DES MATIÈRES** . . . . . **v**

**LISTE DES TABLEAUX** . . . . . **ix**

**LISTE DES FIGURES** . . . . . **x**

**LISTE DES SIGLES** . . . . . **xii**

**REMERCIEMENTS** . . . . . **xiii**

**DÉDICACE** . . . . . **xiv**

**CHAPITRE 1 :INTRODUCTION** . . . . . **1**

1.1 Arbre universel du vivant . . . . . 1

1.1.1 L'arbre des espèces . . . . . 1

1.1.2 Les trois royaumes de la vie . . . . . 1

1.1.3 Concept de procaryote . . . . . 2

1.1.4 Utilisation de caractères moléculaires . . . . . 2

1.2 Histoire et biologie des transferts horizontaux de gènes . . . . . 4

1.2.1 Hérédité et théorie synthétique de l'évolution . . . . . 4

1.2.2 Découverte du transfert horizontal . . . . . 4

1.2.3 Écologie des HGT . . . . . 5

1.2.4 Mécanismes du HGT . . . . . 6

1.3 Fréquence des HGT et remise en cause de l'arbre de la vie . . . . . 14

1.3.1 Fréquence des HGT et type de gènes affectés . . . . . 14

1.3.2 Effet des HGT sur la phylogénie . . . . . 14

1.3.3 Remise en question du concept d'espèce chez les bactéries . . 15

1.3.4 Remise en cause en question de l'utilisation d'un arbre . . . 16
1.3.5 Noyau de gènes peu ou pas transférés . . . . . 17
1.4 Méthodes de détection des HGT et leurs limites . . . . . 17
1.4.1 Méthodes phylogénétiques . . . . . 18
1.4.2 Méthodes non-phylogénétiques . . . . . 21
1.4.3 Limitations des méthodes non phylogénétiques . . . . . 24
1.4.4 Comparaison des méthodes de détection . . . . . 25
1.4.5 Supériorité des méthodes phylogénétiques . . . . . 27
1.5 Phylogénie et HGT . . . . . 27
1.5.1 Hypothèses expliquant la fréquence des HGT . . . . . 27
1.5.2 Notre hypothèse : les gènes répandus sont moins sujets au HGT . . . . . 30
1.5.3 Justification de nos choix méthodologiques . . . . . 30

**CHAPITRE 2 : MATÉRIELS ET MÉTHODES . . . . . 32**

2.1 Protocole . . . . . 32
2.1.1 Données génomiques . . . . . 32
2.1.2 Choix des espèces . . . . . 32
2.1.3 Détection des familles de gènes homologues non-paralogues . 33
2.1.4 Alignement des séquences et sélection des régions bien alignées 37
2.1.5 Reconstruction phylogénétique . . . . . 37
2.1.6 Reconstruction de l'arbre des espèces . . . . . 38
2.2 Extraction des résultats . . . . . 38
2.2.1 Effectifs des familles, des gènes et des groupes testables . . . 38
2.2.2 Comptabilisation des HGT : congruence, incongruence et irrésolution . . . . . 39
2.2.3 HGT artificiels . . . . . 41
2.2.4 Seuil de e-value et arêtes manquantes dans `brh` . . . . . 42
2.2.5 E-value des familles de HNP . . . . . 43
2.2.6 Variation du seuil de significativité du bootstrap . . . . . 44

2.2.7 Effet du nombre d'espèces dans les familles d'HNP . . . . . 44
2.2.8 Catégories de longueur des gènes . . . . . 45
2.2.9 Raccourcissement des gènes . . . . . 45
2.2.10 Conformation des arbres et singletons . . . . . 45
2.2.11 Distribution des gènes chez les espèces avec plusieurs souches 47
2.2.12 Retrait d'espèces . . . . . 48
2.2.13 Tirages aléatoires d'espèces . . . . . 49
2.2.14 Diagramme des programmes . . . . . 49

**CHAPITRE 3 :RÉSULTATS ET DISCUSSION . . . . . 51**

3.1 Exemple d'application de notre protocole : le jeu de données 5x7 . . 51
3.2 HGT artificiels et efficacité du protocole . . . . . 55
3.3 Homogénéisation de la taille des familles : jeu de données 5x2 . . . 57
3.4 Comparaison des méthodes de maximum de vraisemblance et de distance : jeu 5x2 . . . . . 59
3.5 Les résultats irréguliers ne sont pas causés par la taille et la forme des arbres ("Pick from") . . . . . 59
3.6 Impact de la longueur des gènes . . . . . 62
3.7 Influence du seuil de bootstrap . . . . . 64
3.8 Le rééchantillonnage des familles montre que les courbes sont le résultat du signal phylogénétique . . . . . 65
3.9 Influence du seuil d'e-value et du nombre d'arêtes manquantes tolérées sur la détection des familles de HNP par `brh` . . . . . 65
3.10 Classification des familles de HNP en fonction de leur e-value . . . . 68
3.11 Le cas des singletons . . . . . 70
3.12 Absence des gènes souches proches des espèces de référence . . . . . 74
3.13 Application à une échelle évolutive restreinte . . . . . 77
3.14 Autres configurations de l'échantillonnage taxonomique . . . . . 82
3.15 Création de jeux plus petits à partir de 5x7 . . . . . 86
3.16 Tirage aléatoire d'un grand nombre d'espèces . . . . . 88

**CHAPITRE 4 : CONCLUSION ET PERSPECTIVES . . . . . 90**

4.1 Notre hypothèse est vérifiée . . . . . 90

4.2 Évaluation de notre protocole . . . . . 91

4.3 Proportion des gènes détectés par `brh` et leur représentativité . . . 93

4.4 Gènes non détectés par `brh` . . . . . 93

4.5 Amélioration de la détection des HGT . . . . . 94

4.6 Amélioration des analyses . . . . . 95

4.7 Hypothèses alternatives/complémentaires . . . . . 96

4.8 Conclusion . . . . . 96

**BIBLIOGRAPHIE . . . . . 97**

## LISTE DES TABLEAUX


3.1 Seuils e-value pour **brh** . . . . . 66

3.2 Arêtes manquantes pour **brh** . . . . . 67

3.3 Longueurs moyennes des familles avec e-value <1e-40 et >1e-40 . . 68

3.4 Longueurs moyennes des familles pour 5x2 gamma 1e-100 et 1e-4 . 80

4.1 Proportion des gènes détectés par **brh** . . . . . 93

# LISTE DES FIGURES


1.1 Incongruences phylogénétiques causées par les HGT . . . . . 15
1.2 WF Doolittle : l'arbre ou le buisson de la vie ? (Doolittle, 1999) . . 16

2.1 Choix des espèces . . . . . 33
2.2 Détection des familles de gènes HNP . . . . . 35
2.3 Rejet des paralogues . . . . . 36
2.4 Comptabilisation des HGT . . . . . 40
2.5 Simulation d'un HGT . . . . . 42
2.6 Détermination de la e-value pour une famille d'HNP . . . . . 43
2.7 Resampling . . . . . 46
2.8 Pick from . . . . . 46
2.9 Singletons . . . . . 48
2.10 Diagramme des programmes . . . . . 50

3.1 5x7 : arbre . . . . . 52
3.2 5x7 : effectifs . . . . . 53
3.3 5x7 : congruence, incongruence et irrésolution . . . . . 54
3.4 Irrésolution causée par des HGT . . . . . 56
3.5 5x7 : HGT artificiels . . . . . 56
3.6 5x2 original : arbre . . . . . 57
3.7 5x2 original : effectifs . . . . . 58
3.8 5x2 original : congruence, incongruence et irrésolution . . . . . 58
3.9 5x2 original : comparaison ML et distance . . . . . 60
3.10 Pick from . . . . . 60
3.11 Artéfact des arbres de taille 4 . . . . . 61
3.12 Longueur des gènes . . . . . 62
3.13 Raccourcissement des gènes . . . . . 63
3.14 Seuils de bootstrap . . . . . 64
3.15 Rééchantillonnage . . . . . 65

3.16 E-value des familles de HNP . . . . . 69
3.17 Singletons 5x2 original : effectifs . . . . . 71
3.18 Singletons 5x2 original : congruence, incongruence et irrésolution . . 72
3.19 Singletons 5x7 : congruence, incongruence et irrésolution . . . . . 73
3.20 5x2 souches : arbre . . . . . 75
3.21 5x2 souches : comparaison gènes universels vs non-universels . . . . 76
3.22 5x2 gamma : arbre . . . . . 78
3.23 5x2 gamma 1e-4 : congruence, incongruence, irrésolution, effectifs . 78
3.24 5x2 gamma 1e-100 : congruence, incongruence, irrésolution, effectifs 80
3.25 5x2 gamma 1e-4 et 1e-100 : singletons . . . . . 82
3.26 12x2 : arbre . . . . . 83
3.27 12x2 : congruence, incongruence, irrésolution et effectifs . . . . . 84
3.28 6x4 : arbre . . . . . 85
3.29 6x4 : congruence, incongruence, irrésolution et effectifs . . . . . 85
3.30 5x7 : downsampling . . . . . 87
3.31 Tirage de 100 jeux 5x2 : congruence, incongruence et irrésolution ML + MP . . . . . 89

4.1 Fusion de jeux de données . . . . . 92
4.2 Hypothèse de Bravo . . . . . 92

## LISTE DES SIGLES

| | |
|---|---|
| ADN | Acide désoxyribonucléique |
| ALB | Attraction des longues branches |
| ARN | Acide ribonucléique |
| brh | *best reciprocal hit* (meilleur hit BLAST réciproque) |
| `brh` | Le nom de notre programme de détection des brh |
| HGT | *Horizontal gene transfer* (Transfert horizontal de gène) |
| HNP | (gène) Homologue non-paralogue |
| kb | kilo (1000) bases |
| ML | *Maximum likelihood* (Maximum de vraisemblance) |
| MP | *Maximum parsimony* (Maximum de parcimonie) |
| ORF | *Open Reading Frame* |
| pb | paire de bases |
| SNP | Signal non-phyogénétique |
| SP | Signal phyogénétique |
| SSU rRNA | *Small subunit ribosomal RNA* (Petite sous-unité de l'ARN ribosomique) |

# REMERCIEMENTS

Je tiens avant tout à remercier mon directeur **Hervé Philippe** qui à mon humble avis possède les qualités d'un grand scientifique : d'une part la maîtrise parfaite de son domaine et d'excellentes connaissances annexes, associées à une remarquable capacité de synthèse, et d'autre part une ouverture à des idées variées. Ces qualités sont doublées d'une sympathie et d'une modestie qui le rendent très amical et qui créent une amibiance à la fois studieuse et détendue parmi les membres de son équipe.

Je remercie particulièrement **Henner Brinkmann** pour les multiples services rendus, ainsi que les relectures du manuscrit, ainsi que **Denis Baurain**, pour les nombreuses discussions, l'écriture de `forge_brh_datasets`, et l'implémentation des rééchantillonnages dans `interprete`.

Je remercie également les membres actuels ou passés de l'équipe :

Ignacio Bravo[1]
Frédéric Delsuc
Wafae El-Alaoui
Olivier Jeffroy
Claudia Kleinman
Guy Larochelle
Nicolas Lartillot
Boris-Antoine Legault[2]
Nicolas Rodrigue[3]
Naiara Rodriguez Ezpeleta
Béatrice Roure
Yan Zhou

[1] A formulé l'hypothèse de Bravo.
[2] A initialement suggéré que l'irrésolution pouvait aussi dénoter des transferts horizontaux.
[3] A formulé l'hypothèse contraire à la nôtre.

À Marineh ma bien-aimée.

# CHAPITRE 1

# INTRODUCTION

## 1.1 Arbre universel du vivant

### 1.1.1 L'arbre des espèces

Il est communément admis que toutes les formes de vie actuelles descendent d'un ancêtre commun, notamment parce que toutes les formes de vie utilisent (à quelques exceptions mineures près) le même code génétique. Ce concept a été proposé par Darwin (1859) dans son ouvrage *On the origin of species*. Sa théorie comporte trois idées principales : il y a une forte variabilité au sein de chaque population, la sélection naturelle cause la descendance avec modification au cours des générations, et le système naturel de classification est de nature généalogique (Dayrat, 2003). Darwin (1859) représente les relations de parenté entre les espèces sous forme d'arbre. Son schéma est purement théorique car il ne mentionne aucune espèce, mais implique une origine commune pour toutes les espèces. Cependant, il fonde le concept de phylogénie, soit l'origine et l'évolution d'un ensemble d'organismes.

### 1.1.2 Les trois royaumes de la vie

Haeckel (1866) est le premier à proposer une classification globale des différentes formes de vie. L'idée de base est qu'un caractère morphologique semblable chez différentes espèces est une bonne indication d'une origine commune, autrement dit il est improbable qu'il ait évolué plus d'une fois (c'est le concept de parcimonie). Ces états dérivés d'un même caractère ancestral sont dits *homologues*. En comparant plusieurs caractères, on parvient à établir la phylogénie de ces espèces. En utilisant ce principe, Haeckel distingue trois royaumes : *Plantae*, *Animalia* et *Protista* (mais voir Dayrat (2003)). Cependant, la classification des espèces de petite taille est problématique en raison du manque de caractères à comparer. Les espèces microscopiques (et tout ce qui n'est pas reconnu comme animal ou plante)

sont désignées comme *Moneres*, y compris les quelques bactéries déjà découvertes (les "algues bleues-vertes" notamment). Avec le développement d'instruments d'observation plus performants, un plus grand nombre d'espèces microscopiques est découvert, accroissant d'autant l'hétérogénéité de ce royaume.

### 1.1.3 Concept de procaryote

Chatton (1938) est le premier à proposer la dichotomie maintenant bien connue entre procaryotes et eucaryotes, en plus d'introduire les termes mêmes. Cependant ses travaux seront négligés durant des décennies (Sapp, 2005), et ne seront réhabilités qu'en 1962 par Stanier et van Niel. Ces derniers soulignent qu' " il existe deux types d'organisation cellulaire ; les bactéries et les algues bleues-vertes, dont les cellules sont de nature procaryote, se différencient clairement des autres protistes (les autres algues, protozoaires et champignons) qui sont de nature eucaryote" (traduction libre). La différence entre les deux types est essentiellement la présence chez les eucaryotes d'un noyau cellulaire contenant le matériel génétique, ainsi que d'organelles. Malgré ce changement conceptuel radical, déterminer la phylogénie des procaryotes reste une tâche extrêmement ardue. En effet, si les animaux, et les plantes (et même les protistes grâce à l'utilisation du microscope électronique) ont de nombreux caractères morphologiques aidant à les classifier, les procaryotes en présentent relativement peu. Les tentatives de classement doivent recourir essentiellement aux caractéristiques physiologiques des espèces. Ces critères restent très artificiels, et surtout se contredisent souvent. Les résultats obtenus sont loin d'être satisfaisants.

### 1.1.4 Utilisation de caractères moléculaires

Aucun progrès significatif n'a été accompli jusqu'à l'utilisation de caractères moléculaires (Stanier et van Niel, 1962). Zuckerkandl et Pauling (1965), puis Fitch et Margoliash (1967) initient la révolution en proposant les "molécules comme documents de l'histoire évolutive" et leur utilisation pour la construction d'arbres phylo-

génétiques. Woese et Fox (1977) mettent ces concepts en application. Ils utilisent la petite sous-unité (16S ou 18S) de l'ARN ribosomique (SSU rRNA, *small subunit of ribosomal RNA*) pour plusieurs espèces procaryotiques et eucaryotiques. Ils étudient les coefficients d'association (c'est-à-dire le nombre d'oligonucléotides en commun) entre les sous-unités des différentes espèces. Plus les espèces sont proches, plus le nombre d'oligonucléotides identiques entre les molécules sera élevé, et meilleur sera le coefficient d'association. Grâce à cette approche, ils découvrent un groupe de procaryotes (des bactéries méthanogènes et halophiles) qui sont aussi différents des autres bactéries que des eucaroytes. Ils les baptisent *archaebacteria* en référence à leur mode de vie qui semble adapté aux conditions extrêmes qui prévalaient il y a 3 ou 4 milliards d'années. Woese consacre la révolution due à l'utilisation des séquences dans un long plaidoyer (1987). Il expose la supériorité des caractères *génotypiques* sur les caractères *phénotypiques* largement utilisés dans le passé. En effet, la définition et l'interprétation de la similarité des caractères phénotypiques sont souvent subjectives. Par opposition, les caractères génotypiques (c'est-à-dire l'information sur la séquence) sont unidimensionnels. Les éléments d'une séquence nucléotidique ou protéique sont simples, peu nombreux, bien définis, et comparables de façon plus objective en utilisant des approches mathématiques.

Ainsi, le mécanisme sous-jacent à l'évolution est mis en lumière : il s'agit des mutations continuelles au niveau de la séquence génomique, souvent sans effet, mais qui résultent parfois en un changement du phénotype. C'est ce phénotype qui, en fonction de sa valeur d'adaptation (*fitness*[1]), sera favorisé ou non par la pression sélective.

L'utilisation de caractères moléculaires a constitué une révolution dans le domaine de la phylogénie, permettant des analyses beaucoup plus rigoureuses qu'avant. Cela a également permis l'unification des classifications des espèces, notamment chez les "procaryotes", représentés en fait par deux royaumes bien distincts, les Bactéries et les Archées. Toutes les autres formes de vie connues jusqu'alors constituent les Eucaryotes (à l'exception peut-être des virus). Le nom de ces trois *domaines* de

[1]Mesure du succès reproductif.

la vie (*Eucarya*, *Bacteria* et *Archaea*) a été proposé par Woese et al. (1990).

## 1.2 Histoire et biologie des transferts horizontaux de gènes

### 1.2.1 Hérédité et théorie synthétique de l'évolution

Afin d'expliquer les mécanismes de l'hérédité, Darwin (1868; 1871) a proposé la théorie de la pangenèse. Selon cette théorie, toutes les cellules du corps contribuent à la formation des gamètes en envoyant des pangènes vers les organes reproducteurs. Cette théorie s'est avérée erronée. C'est Mendel (1866) qui avec ses expérimentations sur les pois énonce les trois lois de l'hérédité qui sont tenues pour vraies aujourd'hui encore. Ses travaux passent relativement inaperçus, et c'est seulement en 1900 qu'ils sont redécouverts de façon indépendante et simultanée par de Vries, Carl Correns et von Tschermak. Par la suite, grâce à leurs travaux sur *Drosophilia melanogaster*, Morgan et al. (1915) montrent que le support de l'hérédité sont des gènes situés de façon linéaire sur des chromosomes. Cependant, ce mode d'évolution est discontinu, et les tenants de l'approche biostatistique (notamment Pearson et Weldon) s'y opposaient vigoureusement, arguant que l'évolution était un processus continu. Fisher (1918) propose un modèle statistique pour l'hérédité mendélienne qui réconcilie les deux conceptions en montrant comment une variation continue peut être le résultat de l'action de plusieurs loci discrets. Ce modèle qui unifie les approches continues et discrètes pour expliquer l'évolution constitue la base de la théorie synthéthique de l'évolution. Celle-ci est appuyée par d'autres (Dobzhansky, 1937; Mayr, 1942) et gagne rapidement la faveur de la communauté scientifique. Le dogme central est la transmission d'une génération à l'autre du matériel génétique. Cependant, cette transmission verticale ne pouvait pas expliquer certains phénomènes constatés chez les bactéries.

### 1.2.2 Découverte du transfert horizontal

Il avait été remarqué que des phénotypes pouvaient se répandre parmi des cultures bactériennes, suggérant l'échange de matériel génétique entre individus,

et non pas seulement entre cellules mères et filles. Comme le résument Thomas et Nielsen (2005), la preuve de l'existence de ce type de transfert a été apportée par la découverte que la virulence était transmissible entre pneumocoques chez la souris infectée (Griffith, 1928). Par la suite, Avery et al. (1944) ont découvert le principe transformant responsable : ayant isolé le "sodium desoxyribonucleate" (l'ADN) de pneumocoques de type III (avec une capsule de polysaccharides), ils ont réussi à transformer des pneumocoques de type II en ceux de type III uniquement à partir de l'ADN de type III. De plus, ils remarquent que cette modification est "prédictible, spécifique et transmissible (*heritable*)" (traduction libre). Cette expérience historique démontre que le support de l'hérédité est la molécule d'ADN. Elle met aussi en lumière l'échange de matériel génétique entre individus, baptisé transfert *horizontal* (HGT, pour *horizontal gene transfer*) par opposition à la transmission héréditaire, dite *verticale*. Par la suite, plusieurs autres cas de transfert horizontal ont été mis à jour, permettant une meilleure compréhension de ce phénomène.

### 1.2.3 Écologie des HGT

**Îlots génomiques** La découverte de régions génomiques aux caractéristiques particulières a conduit à suspecter une occurence de HGT. Au début des années 1980, Hacker et al. (1983) observent chez *E. coli* des régions dont le taux de G+C et l'utilisation des codons diffèrent du reste du génome. Ces régions, nommées *îlots génomiques*, codent pour des fonctions non-essentielles comme la résistance aux antibiotiques, des propriétés impliquées dans la symbiose ou la pathogenèse, ainsi que d'autres fonctions métaboliques (voir Hacker et Kaper (2000)). Hacker et al. (1983) remarquent que les gènes associés à la virulence diffèrent d'une souche d'*E. coli* à une autre. Ces régions, qui sont alors baptisées *îlots de pathogénécité* (Hacker et al., 1990), ont potentiellement été acquises par HGT. Plusieurs éléments soutiennent cette hypothèse : la présence de répétitions directes à leurs extrémités, la présence de déterminants d'intégrases et autres loci de mobilité, ainsi que leur instabilité génétique.

**HGT et caractéristiques écologiques** Les HGT semblent aussi être influencés par les caractéristiques écologiques des espèces. Jain et al. (2003) ont mis en évidence une corrélation positive entre la fréquence des HGT et la taille du génome et le contenu en G+C du génome. L'utilisation du carbone, le taux d'oxygène et les températures de croissance similaires sont moyennement corrélés. Les HGT se font donc préférentiellement entre espèces partageant des caractéristiques similaires. L'argument pourrait paraître circulaire, mais les deux facteurs les plus importants (taille du génome et son contenu en G+C) sont indépendants de la proximité physique entre les espèces.

Coleman et al. (2006) ont étudié les réactions physiologiques face aux variations de lumière et de nutriments de plusieurs souches de *Prochlorococcus*. Les souches présentent une grande variabilité génomique leur permettant de s'adapter aux conditions prévalant dans leur habitat. L'essentiel des différences est concentré dans les îlots génomiques. Les caractéristiques de ces régions indiquent qu'elles ont vraisemblablement été acquises par HGT : elles sont associées à des gènes d'ARNt (qui sont un site courant d'intégration pour les éléments mobiles), l'ARNt-proline est répété plusieurs fois aux extrémités des îlots, et surtout, jusqu'à 80% des gènes de ces régions sont plus similaires à ceux d'organismes hors des cyanobactéries (auxquelles appartiennent *Prochlorococcus*). Le nombre et la variété des fonctions remplies par les gènes dans les îlots génomiques mettent en évidence l'importance du rôle des HGT dans l'adaptation d'espèces bactériennes à de nouvelles niches écologiques.

### 1.2.4 Mécanismes du HGT

Le transfert horizontal de gènes comporte trois étapes principales : le transfert du matériel génétique dans le cytoplasme de la cellule hôte, l'intégration dans le génome de la cellule hôte et la fixation dans la population.

#### 1.2.4.1 Transfert

Le transfert proprement dit peut s'accomplir de trois façons : la transformation naturelle, la conjugaison et la transduction.

**1.2.4.1.1 Transformation naturelle** La transformation est l'acquisition de matériel génétique présent dans l'environnement et son intégration dans le génome.

**ADN dans l'environnement** Contrairement à ce que l'on pourrait concevoir intuitivement, l'ADN est présent dans l'environnement : il provient de cellules détruites ou en décomposition. De plus, certains genres le sécrètent naturellement : *Acinetobacter*, *Alcaligenes*, *Azotobacter*, *Bacillus*, *Flavobacterium*, *Micrococcus*, *Pseudomonas* et *Streptococcus* (Lorenz et Wackernagel, 1994; Moscoso et Claverys, 2004; Paget et Simonet, 1994). Des quantités non négligeables d'ADN libre sont présentes : 1 $\mu$g/g dans le sol et les sédiments (Ogram et al., 1987), et de 0,03 à 88 $\mu$g dans l'eau douce et de mer (DeFlaun et Paul, 1989; Karl et Bailiff, 1989). De plus, de nombreuses expériences ont montré que l'ADN libéré dans divers environnements n'est pas immédiatement dégradé, mais persiste pour des durées allant de quelques heures à quelques jours (voir Thomas et Nielsen (2005)).

**Compétence** Les cellules bactériennes doivent entrer dans un état physiologique régulé nommé *compétence* qui implique 20 à 50 protéines. Les espèces bactériennes naturellement transformables développent l'état de compétence sous des conditions particulières comme une modification de l'environnement, l'accès aux nutriments, ou la densité cellulaire (Thomas et Nielsen, 2005). On estime qu'environ 1% des espèces décrites sont naturellement transformables (Jonas et al., 2001), mais représentent une grande variété taxonomique : chez les Archaea ainsi que dans les grands groupes bactériens (Lorenz et Wackernagel, 1994; Paget et Simonet, 1994)
.

**Entrée de l'ADN** L'ADN se lie de façon non covalente à des récepteurs spécifiques à la surface de cellules compétentes. Dans la plupart des cas, la liaison est non spécifique, mais *N. gonorrhoeae* et *H. influenzae* par exemple reconnaissent des régions de 9 à 11 pb. Celles-ci sont espacées de 4 à 5 kb dans leur propre génome, ce qui favorise les HGT intra-spécifiques. Dans tous les cas, l'ADN traverse la membrane sous forme simple brin (Thomas et Nielsen, 2005).

**1.2.4.1.2 Conjugaison** La conjugaison est le transfert de matériel génétique entre deux cellules via une jonction nommée *pilus*.

**Plasmides et ICE** Le matériel transféré est le plus souvent des *plasmides* ou des *ICE* (*integrative conjugative element*). Un plasmide est un élément génétique stable, circulaire (il existe aussi des plasmides linéaires), à double brin et auto-réplicatif (Frost et al., 2005; Thomas et Nielsen, 2005). Il contient des gènes pour leur propre réplication ainsi que les gènes pour leur transfert (Frost et al., 2005). La conjugaison est la voie privilégiée pour la transmission de la résistance aux médicaments. Les ICE sont des groupes de gènes chromosomiques qui codent pour des intégrases et des protéines de conjugaison, ainsi que d'autres fonctions. De plus, les ICE et les plasmides peuvent être transférés entre les cellules, contrairement aux îlots génomiques (Frost et al., 2005). La conjugaison comporte trois étapes : l'appariement des cellules et la formation du pilus, la signalisation que le transfert peut débuter, et le transfert proprement dit. Il existe différents types de mécanismes de conjugaison, les deux plus distincts étant ceux des bactéries Gram-positives et Gram-négatives (Frost et al., 2005).

**Appariement des cellules** Un gène code pour la "protéine de couplage" qui est responsable de la synchronisation de l'appariement des cellules, ainsi que du "pompage" de l'ADN dans la cellule receveuse (Gomis-Ruth et al., 2004). Il s'agit d'une protéine de la famille des ATPases TraG qui se lie à la membrane cytoplasmique. La plupart des systèmes de conjugaison comprennent une relaxase

qui transforme l'ADN en simple brin pour en faciliter le transfert (Frost et al., 2005).

**Formation du pilus** Le pilus est assemblé par le système de sécrétion de type IV (T4SS) dans lequel une protéine de couplage relie un complexe protéique transmembranaire (un *transférosome*) à un complexe nucléoprotéique (un *relaxosome*). Les T4SS se retrouvent dans tous les plasmides (sauf pour *Bacteroides*) ainsi que dans plusieurs ICE (voir (Frost et al., 2005)).

**Transfert de l'ADN** Il est intéressant de souligner que les plasmides et les ICE peuvent aussi transférer de l'ADN chromosomique (Wilkins et Frost, 2001). En effet, ils peuvent parfois être intégrés dans le génome hôte, et sont alors excisés de façon imprécise avec des portions du génome hôte potentiellement grandes (Frost et al., 2005). Lorsque le plasmide ou l'ICE passent dans la cellule réceptrice, l'ADN de l'hôte se retrouve transféré horizontalement.

**1.2.4.1.3 Transduction** La transduction est un transfert de gènes médié par certains phages.

**Bactériophages** Les bactériophages (ou phages) sont des virus s'attaquant aux bactéries. Leur génome peut être constitué d'ADN ou d'ARN simple- ou double-brin de taille de 5 à 650 kilobases (kb). Leur mode de vie consiste à détourner les systèmes de la cellule-hôte afin de lui faire synthétiser de nouveaux virus (protéines de la capside et génome). Il existe deux types de bactériophages : les virulents qui causent la lyse de la cellule infectée, et les tempérés, qui ont un cycle alternatif au cours duquel ils s'intègrent au génome-hôte. Cette recombinaison non-homologue est médiée par des enzymes encodées par leur génome (Frost et al., 2005).

**Transfert de l'ADN** Lorsque le virus entre dans le cycle lytique, l'ADN de l'hôte est parfois encapsidé avec le génome viral (Zinder et Lederberg, 1952). Lors de l'éclatement de la cellule, les nouveaux virus se retrouvent dans l'environnement.

Après qu'ils ont infecté d'autres cellules, l'ADN de l'hôte original est intégré dans le génome du nouvel hôte (et ainsi transféré horizontalement).

#### 1.2.4.2 Intégration dans le génome

Une fois que l'ADN étranger est entré dans la cellule, il doit être intégré au génome-hôte afin de pouvoir persister de génération en génération. La séquence transférée peut être intégrée par recombinaison homologue si elle présente une similarité suffisante avec une partie du génôme hôte. Dans le cas contraire, l'intégration peut quand même se faire par recombinaison illégitime. Les plasmides constituent toutefois une exception : ces molécules d'ADN généralement circulaires et double-brin possèdent une origine de réplication et sont capables de s'autorépliquer en utilisant la machinerie de la cellule-hôte (Thomas et Nielsen, 2005), et n'ont donc pas besoin d'être intégrés pour persister.

**1.2.4.2.1 Recombinaison homologue** Une fois dans la cellule, l'ADN peut s'intégrer au génome hôte par *recombinaison homologue* (après être redevenu double-brin au préalable). Pour cela l'ADN et le génome hôte doivent présenter une région de 25 à 200 paires de bases (pb) avec une grande similarité, c'est-à-dire au plus 25% de différence (Thomas et Nielsen, 2005). Le taux de recombinaison varie de 0,1% (*A. babyli*) à 25-50% (*B. subtilis* et *S. pneumoniae*) (Palmen et Hellingwerf, 1997). Gomez et al. (2005) ont comparé la fréquence de recombinaison de plasmides thêta chez *E. coli* et *B. subtilis* en fonction du pourcentage de divergence des séquences : chez *E. coli* elle passe de $10^{-4}$ pour 0% de divergence, à $10^{-8}$ pour 4% et 22% de divergence. Chez *B. subtilis*, elle varie de $10^{-3}$ pour 0% à $10^{-4}$ pour 22% de divergence.

La recombinaison homologue se produira donc préférentiellement entre gènes issus de la même espèce ou d'espèces proches, et surtout entre souches au sein d'une même espèce (Ochman et al. (2005)). Cependant, il existe des portions de gènes très conservées à cause d'une forte pression de sélection comme l'ARNr, et des gènes homologues d'espèces très éloignées pourront être recombinées. La

recombinaison homologue est un phénomène rare (fréquence de $10^{-3}$ dans le cas optimal de séquences identiques) et difficile à détecter car ses effets sont difficilement discernables de substitutions aléatoires. Heureusement, ce phénomène ne semble pas trop affecter l'inférence phylogénétique (Philippe et Douady, 2003), du moins dans un contexte de concaténation phylogénomique. Au vu de la rareté de ce phénomène, nous pouvons nous permettre de ne pas nous en soucier.

**1.2.4.2.2 Recombinaison illégitime** Plusieurs gènes permettent la *recombinaison illégitime* chez *E. coli* (Ikeda et al., 2004). Ce type de recombinaison nécessite peu (9-13 pb) ou très peu (<4 pb) d'identité entre les séquences (van Rijk et Bloemendal, 2003), et semble dépendre de la quantité d'ADN ligase présente : présente en quantité suffisante, cette enzyme permet l'intégration illégitime de fragments d'ADN normalement trop courts pour être intégrés (Onda et al., 2001). D'après Thomas et Nielsen (2005), ce phénomène serait relativement rare, mais selon van Rijk et Bloemendal (2003) la recombinaison illégitime est utilisée par les cellules de mammifères pour réparer les cassures double-brin dans l'ADN, et ce 100 à 1000 fois plus fréquemment que la recombinaison homologue.

### 1.2.4.3 Fixation dans la population

Une séquence d'ADN transférée horizontalement l'est dans une seule cellule. Afin de demeurer de façon pérenne chez l'espèce réceptrice, elle doit se répandre chez tous les individus. Lorsque cela s'est produit, on dit qu'il y a eu *fixation* (la définition classique de la fixation est le moment où la fréquence d'un allèle donné atteint 1, soit 100%).

**Facteurs s'opposant à la fixation** Berg et Kurland (2002) ont étudié et modélisé la dynamique de fixation des HGT au sein de populations bactériennes. Un grand nombre de facteurs concourent à empêcher la fixation. Premièrement, on s'attend à ce que la très grande majorité des séquences transférées n'apportent pas de nouvelle fonction ou d'avantage sélectif à leur hôte. C'est dire que la plupart du

temps, les séquences transférées sont **neutres** ou presque neutres (*"nearly-neutral"*) (Berg et Kurland, 2002). D'autre part, la probabilité qu'une séquence n'ayant pas d'avantage sélectif soit fixée est inversement proportionelle à la taille de la population $N$. Plus précisément, elle est de $1/N$ (Kimura, 1962). Ainsi, pour les populations bactériennes, dont les tailles sont très grandes (la population globale d'*E. coli* a été estimée à $10^{20}$ (Ochman et Wilson, 1987)), la **probabilité de fixation** d'une séquence neutre est presque nulle. Il faut aussi tenir compte de l'effet des **mutations aléatoires** : celles-ci s'opposent à la fixation à cause de leurs effets délétaires. Enfin, le **coût de maintien** du matériel génétique supplémentaire à entretenir cause une perte de compétitivité (*"loss of fitness"*) pour l'hôte. À terme, ces phénomènes causent la disparition pure et simple des séquences transférées. Le processus se nomme **nettoyage génétique** (*"genetic sweep"*). La durée de persistence des séquences transférées dans les génomes est donc faible à l'échelle évolutive. D'autres effets préviennent la fixation. Une séquence importée sera **moins efficace** chez l'hôte car elle a été optimisée pour sa traduction et ses interactions dans l'environnement créé par le génome d'origine. Dans le cas où le gène importé a déjà un homologue dans le génome hôte, la probabilité qu'il soit fixé en remplacement de l'original est égale à $1/2N$ (Berg et Kurland, 2002), ce qui est presque égal à zéro pour les populations bactériennes. Par conséquent, un gène redondant aura tendance à acquérir une nouvelle fonction par mutation plutôt que de conserver sa fonction originale dans le cas où il aura été fixé dans la population.

**Condition de fixation** Pour être fixé, un gène transféré doit donc apporter un *avantage sélectif* suffisamment important pour contrebalancer tous les facteurs concourant à son élimination (Berg et Kurland, 2002). On imagine aisément que la fréquence de fixation d'une séquence transférée est presque nulle à l'échelle courante. Autrement dit, un HGT réussi semble un évènement rare, à moins qu'on se place à l'échelle évolutive. Cependant, les travaux de Hao et Golding (2004) suggèrent qu'un grand nombre de transferts horizontaux sont présents aux extrémités de l'arbre de la vie, impliquant donc les espèces vivant actuellement. Certains de

ces gènes ne sont pas présents chez toutes les souches des espèces étudiées, et serviraient à l'adaptation écologique, ce qui est confirmé par Marri et al. (2006). Hao et Golding (2006) indiquent que ces gènes acquis sont lignée-spécifiques et qu'ils évoluent plus rapidement que les gènes résidants. Cette évolution rapide serait un mélange de sélection directionnelle (*directional selection*) pour s'harmoniser avec les caractéristiques du génome, et de mutations détruisant leur fonction. Ces gènes seraient donc acquis et sélectionnés rapidement pour l'adaptation à une niche écologique particulière, et vraisemblablement perdus rapidement lors d'un changement de niche de l'espèce.

D'autre part, le fait qu'une séquence doive être sélectivement avantageuse pour être fixée (tel que prédit par le modèle de Berg et Kurland (2002)) est contesté par Novozhilov et al. (2005). Leur modélisation mathématique des HGT, qui est une extension du modèle de Berg et Kurland (2002), permet à des séquences neutres ou même légèrement désavantageuses d'être conservées durant d'assez longues périodes. Le modèle comporte cinq paramètres : le taux d'inactivation par mutation, le coefficient de sélection, le taux d'invasion (le taux d'arrivée de nouvelles séquences dans la population), le taux de propagation intra-population ("infection"), et la taille de la population. Si le taux de propagation intra-population est comparable au taux d'inactivation par mutation, et en considérant le processus d'"invasion" (deux paramètres que Berg et Kurland (2002) ont négligés), des séquences neutres ou légèrement désavantageuses pourront persister assez longtemps pour être fixées. Malheureusement, il n'existe pas d'estimation empirique de ces paramètres, ce qui empêche de vérifier la validité du modèle.

Le transfert horizontal de gènes est donc un processus complexe pouvant s'accomplir de façons très variées, allant d'un transfert direct de matériel génétique entre deux cellules bactériennes à une transmission via un médium viral. Si le transfert horizontal permet l'acquisition rapide de nouvelles fonctions permettant de s'adapter à une nouvelle niche ou à un changement écologique, sa réussite est contrecarrée par un grand nombre de facteurs adverses. Un gène transféré doit

notamment interagir avec de nouveaux partenaires génomiques, et doit aussi se répandre parmi tous les individus de la population pour être fixé et ainsi perpétué.

## 1.3 Fréquence des HGT et remise en cause de l'arbre de la vie

### 1.3.1 Fréquence des HGT et type de gènes affectés

La fréquence des transferts horizontaux de gènes est difficile à estimer. Initialement, peu d'exemples étaient décrits, et l'idée générale étaient que les transferts étaient des évènements plutôt rares. Sous l'influence de Woese et al. (1980) et de RF Doolittle (Smith et al., 1992), l'hypothèse d'un HGT ne devait être appelée qu'en dernier recours, quand toutes les autres explications avaient échoué. Cela a mené au sentiment que les HGT étaient rares. Cependant, avec le début de l'ère génomique et la disponibilité des séquences de génomes entiers, un grand nombre de cas ont été découverts. Nakamura et al. (2004) ont par exemple estimé à 14% le nombre de gènes acquis récemment par HGT dans les génomes de 116 procaryotes.

Gogarten et al. (2002) recensent plusieurs exemples de HGT impliquant l'ARN ribosomique, des protéines ribosomiques, des facteurs d'élongation, etc., ainsi que plusieurs enzymes métaboliques comme l'A/F-ATPase, la glutamate synthase, une catalase/péroxydase, etc. Une grande variété de groupes taxonomiques sont impliqués. Deux bactéries hyperthermophiles, *Aquifex aeolicus* et *Thermotoga maritima*, ont un grand nombre de gènes d'origine clairement archaenne (Aravind et al., 2001; Nelson et al., 1999). Le génome de *Synechocystis sp.* code pour un certain nombre de protéines associées à différentes formes de signalisation d'origine eucaryotique (Kaneko et Tabata, 1997; Ponting et al., 1999). Ces exemples illustrent l'existence de transferts inter-domaines.

### 1.3.2 Effet des HGT sur la phylogénie

En plus de leurs effets marqués sur les organismes, les HGT peuvent influencer l'inférence phylogénétique de façon drastique. En effet, si deux espèces éloignées sont impliquées dans un transfert, alors la phylogénie du gène transféré les fera

paraître beaucoup plus proches qu'elles ne le sont en réalité. Un tel réarrangement de la phylogénie peut la rendre totalement incongruente avec la phylogénie des espèces (Philippe et Douady, 2003) ; voir figure 1.1.

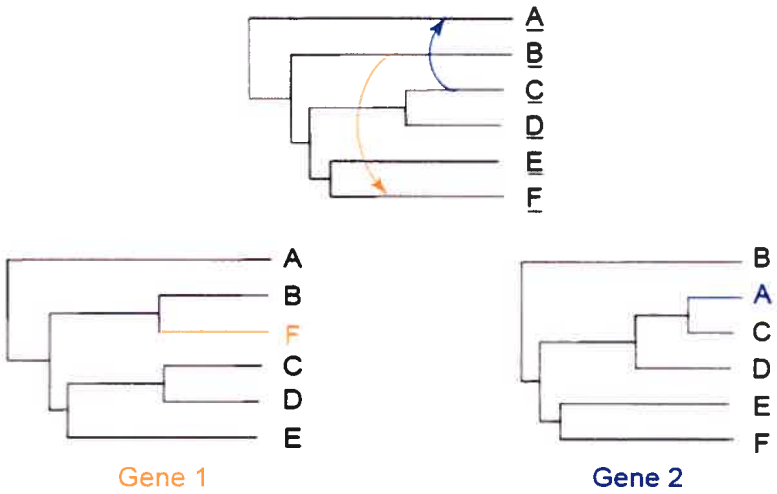


Figure 1.1 – Incongruences phylogénétiques causées par les HGT (Philippe et Douady, 2003). Les flèches orange et bleue dénotent deux HGT hypothétiques affectant les gènes 1 et 2, respectivement. Les phylogénies résultantes sont complètement incongruentes, bien que le nombre de HGT soit faible.

Si un seul transfert récent peut être difficile à détecter, comment démêler le grand nombre de transferts recensés ? Cependant, il convient de mentionner que toutes les incongruences ne sont pas causées par les HGT (Philippe et Douady, 2003). En effet, les artéfacts de reconstruction en sont une source majeure (voir section 1.4.1). De plus, l'ancienneté du transfert peut également avoir un impact important : par exemple un transfert pourrait être si ancien (et donc très répandu dans les espèces actuelles) qu'on ne soupçonnerait pas qu'il a eu lieu (Lawrence et Ochman, 2002). Il apparaît donc que la phylogénie des gènes peut être dissociée de la phylogénie des organismes (la phylogénie traditionnelle).

### 1.3.3 Remise en question du concept d'espèce chez les bactéries

Si l'on extrapole la dissociation entre phylogénie des gènes et phylogénie des espèces, on peut aller jusqu'à redéfinir la notion même d'espèce. Chez les proca-

ryotes en général – et chez les bactéries en particulier – où les taux de HGT sont potentiellement élevés, une espèce peut être redéfinie comme un réservoir duquel rentrent et sortent les gènes au gré des transferts (Gogarten et Townsend, 2005). Les procaryotes formeraient alors une communauté partageant un bassin de gènes, chaque espèce se définissant par le sous-ensemble de gènes possédés. Si certaines espèces échangent des gènes de façon préférentielle entre elles, la fréquence de transfert pourrait devenir le signal phylogénétique (Gogarten et al., 2002). Dans ce cas, deux espèces seraient plus proches d'une troisième non pas parce qu'elles ont un ancêtre commun plus récent, mais parce qu'elle s'échangent des gènes plus fréquemment qu'avec la troisième.

### 1.3.4 Remise en cause en question de l'utilisation d'un arbre

Sans aller jusqu'à une conception si radicale, la multiplication des cas de HGT reportés a conduit WF Doolittle (1999) à remettre en cause le concept d'arbre de la vie. En effet, si chaque transfert est représenté par une ligne horizontale entre les branches impliquées, l'arbre de la vie ressemblerait plus à un réseau.

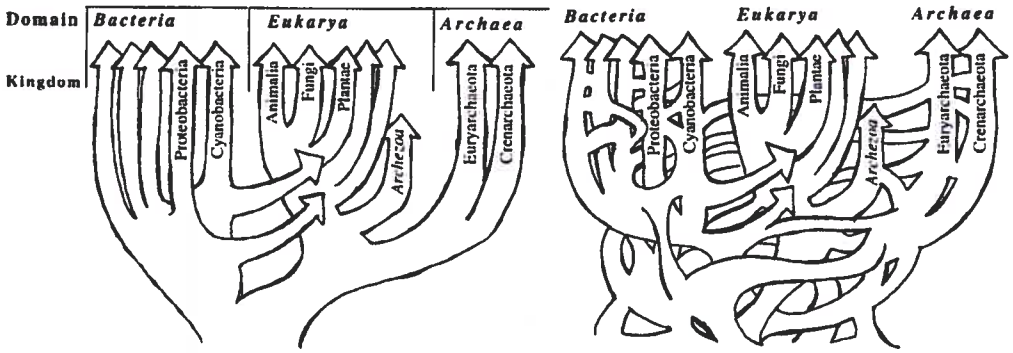


Figure 1.2 – WF Doolittle : l'arbre ou le buisson de la vie ? (Doolittle, 1999)

Des auteurs ont proposé des concepts moins contraignant qu'un arbre, comme par exemple une "toile d'araignée" (*cobweb*) de la vie (Ge et al., 2005), où le tronc et les branches de l'arbre (représentant les relations de parenté) sont plus épais, et où les transferts sont représentés par des filaments entre les branches. Bapteste et al. (2004) proposent la "Synthèse de la Vie" (*The Synthesis of Life*), qui modélise

les deux composantes (verticale et horizontale) de l'hérédité. Il s'agit d'un arbre phylogénétique classique dont certains noeuds sont reliés de façon horizontale afin de représenter les HGT. Selon eux, les portions bien résolues représentent un signal vertical prédominant. Les portions incongruentes et moins bien résolues dénotent des gènes ayant récemment subi des évènements de HGT. Toutes ces propositions sont globalement équivalentes car elles peuvent être résumées par un réseau. Cette représentation est plus souple que l'arbre et permet à une espèce d'avoir plusieurs ancêtres (modélisation du HGT). Cependant, comme la plupart des algorithmes relatifs à la phylogénie ont été développés pour les arbres, les réseaux sont encore relativement peu utilisés. De plus, les réseaux étant plus complexes, les temps d'exécution des algorithmes sont plus grands.

### 1.3.5 Noyau de gènes peu ou pas transférés

Évidemment, il serait un peu naïf de sombrer dans le "tout HGT". L'existence d'un noyau de gènes résistants aux transferts semble intuitive (c'est la conception traditionnelle de l'hérédité). Plusieurs études tendent à montrer qu'un tel noyau existe (Snel et al., 2002). Woese et al. (1980) ont montré que la petite sous-unité du rRNA était une bonne molécule pour construire des phylogénies car elle est présente chez tous les organismes. Cela devrait rendre ce gène résistant au HGT, ce qui est presque vrai : seuls deux transferts ont été documentés (Kurland et al., 2003).

## 1.4 Méthodes de détection des HGT et leurs limites

Nous avons discuté de la nature des HGT, leur mécanisme, leur fréquence. Cependant, nous n'avons pas abordé la délicate question de leur mise en évidence, ce que nous allons faire maintenant. Il existe deux grandes catégories de méthodes permettant la détection des HGT, les phylogénétiques et les non-phylogénétiques ou alternatives. Les méthodes phylogénétiques requièrent la contruction d'arbres pour chaque gène, puis la détection de discordances avec l'arbre des espèces. Les mé-

thodes non-phylogénétiques utilisent les caractéristiques génomiques pour détecter les régions qui semblent anormales, et qui ont donc potentiellement été transférées.

### 1.4.1 Méthodes phylogénétiques

La méthode phylogénétique est la méthode traditionnelle pour détecter les HGT. C'est aussi la façon la plus naturelle car il s'agit de retrouver dans l'information verticale (ou héréditaire ; c'est la définition de la phylogénie) les incohérences dues à une évolution non-verticale (c'est-à-dire horizontale). Lors de la construction d'arbres phylogénétiques pour différents gènes putativement orthologues, on obtient parfois des résultats incongruents. Ces incongruences ont néanmoins plusieurs explications : les artéfacts de reconstruction, la perte de gènes, la paralogie et les HGT. Pour des revues sur les artéfacts de construction, se reporter à Sanderson et Shaffer (2003) et Philippe et al. (2005).

#### Artéfacts de reconstruction

Plusieurs artéfacts de reconstruction influencent les résultats obtenus. Par exemple, Felsentstein (1978) a décrit l'attraction des longues branches ($ALB$) : les espèces évoluant rapidement (donc représentées par des branches plus longues) sont regroupées ou attirées même si elles n'ont pas de relation de parenté récente, quand on infère l'arbre par maximum de parcimonie. Le problème est particulièrement aigu lorsque l'arbre considéré comporte de longues branches à la base (par exemple l'arbre de la vie avec les trois domaines, ou l'ensemble des espèces d'un domaine). Dans ce cas, les espèces évoluant rapidement semblent émerger plus près de la racine de l'arbre. Un exemple célèbre est le branchement artéfactuel des microsporidies dans les arbres de la vie initialement proposés par Vossbrinck et Woese (1986) ; voir aussi Brinkmann et al. (2005).

Un autre exemple d'artéfact est lié au biais de composition. Les séquences ayant des taux de G+C similaires sont attirées artéfactuellement dans l'arbre (Woese et al., 1991), de sorte que des espèces très éloignées mais avec des taux de G+C

égaux peuvent être groupées (voir Lockhart et al. (1994)). De même, il existe une attraction pour les espèces ayant des tailles de génomes similaires quand on construit l'arbre en utilisant le contenu en gènes (Lake et Rivera, 2004).

### Paralogie et perte de gènes

La duplication de gènes donne naissance à des familles de paralogues qui évoluent et acquièrent des fonctions différentes. Parfois, seuls certains exemplaires des paralogues sont perdus. Il arrive fréquemment que certaines lignées perdent des gènes, conduisant à une distribution en apparence anormale des orthologues restants (Lawrence et Ochman, 2002). Lors de la reconstruction phylogénétique, les paralogues peuvent être aisément confondus avec des orthologues : si l'on considère un groupe d'orthologues, mais que l'un d'eux est en réalité un paralogue, ce gène paraîtra phylogénétiquement plus éloigné des autres de façon artéfactuelle. On obtient ainsi une phylogénie erronnée. Ce phénomène de perte différentielle de paralogues correspond à ce que l'on appelle la paralogie cachée.

### HGT

Lorsque les artéfacts de reconstruction et la paralogie cachée sont exclus, l'explication la plus vraisemblable des incongruences est le HGT. Les méthodes phylogénétiques consistent à établir une phylogénie de référence pour les espèces, généralement avec un ou plusieurs gènes universels (par exemple la SSU rRNA), puis à comparer les phylogénies obtenues avec différents gènes et identifier les incongruences (Brochier et al., 2002; Daubin et al., 2003b).

### Détection des "orthologues"

**Homologues non-paralogues** La première étape des méthodes phylogénétiques est fondée sur la détection de gènes "orthologues", ou plus exactement des

gènes homologues non-paralogues (HNP)[2]. Les gènes orthologues sont le résultat des spéciations : ils représentent donc la phylogénie des espèces. Au contraire, les xénologues étant le produit de transferts horizontaux, une phylogénie les contenant ne reflètera pas la phylogénie des espèces. C'est donc la recherche d'incongruences parmi les phylogénies obtenues avec des HNP qui permet la détection de transferts.

**Principe de recherche** Comme les HNP ont une origine commune, on s'attend à ce qu'ils présentent une certaine similarité au niveau de leur séquence. Le programme BLAST (Altschul et al., 1997) permet la recherche de telles similarités.

**Clustering** Mushegian et Koonin (1996) introduisent le concept de *best reciprocal hit* (brh) : soient $A$ et $B$ des ensembles disjoints de gènes. Soit $\alpha$ un gène appartenant à $A$. Soit $\beta$ le meilleur hit de $\alpha$ dans $B$ (obtenu en "blastant" $\alpha$ contre $B$, c'est-à-dire en utilisant BLAST à partir de $\alpha$ afin de trouver les gènes les plus similaires dans $B$). Si le meilleur hit de $\beta$ dans $A$ est $\alpha$, alors $\alpha$ et $\beta$ constituent un *best reciprocal hit.* Ainsi, $\alpha$ et $\beta$ étant les gènes les plus similaires dans les deux génomes, ils sont vraisemblablement orthologues.

Tatusov et al. (1997) utilisent les *best hits* (non réciproques) entre plusieurs génomes pour en extraire les groupes d'orthologues (COG, *Clusters of Orthologous Groups*). Partant d'un groupe de trois gènes ayant des *best hits* symétriques (des brh) formant donc une relation triangulaire, ils étendent le COG avec des *best hits* parmi d'autres organismes, toujours avec des *best hits* triangulaires. La méthode des brh a été utilisée par un grand nombre de chercheurs (voir par exemple Ragan et Charlebois (2002); Zhaxybayeva et al. (2006)), et a servi d'inspiration à d'autres : Fulton et al. (2006) ont créé une méthode nommée Ortholuge. Cette méthode détecte les orthologues putatifs par brh, puis exclut les paralogues potentiels. Ceci est réalisé en comparant le rapport de la distance phylogénétique entre deux espèces et un outgroup à la fois dans l'arbre du gène et dans l'arbre des espèces : si ce

[2]En effet, les gènes acquis par transfert horizontal ne sont pas orthologues, mais xénologues (Gray et Fitch, 1983). Cependant, le terme "orthologue" est couramment utilisé dans la littérature. Nous nous intéressons donc aux HNP (orthologues et xénologues) pour la détection des HGT.

rapport est trop différent dans les deux arbres, alors le gène détecté est en fait un paralogue.

Wall et al. (2003) ont créé une méthode reprenant le concept de réciprocité nommée *rsd* (*reciprocal smallest distance*). Cet algorithme utilise l'alignement global des séquences et l'estimation par maximum de vraisemblance des distances évolutives pour détecter les orthologues entre deux génomes. Selon les auteurs, cette méthode est moins sensible aux paralogues que la méthode brh.

Beiko et al. (2005) utilisent l'algorithme de clusterisation Markovienne hybride de Harlow et al. (2004) afin d'identifier les familles de gènes. Cette méthode se base sur les séquences, et combine les avantages de la clusterisation Markovienne (évite les clusters avec les domaines similaires) et de la clusterisation simple-lien (*single-linkage*) qui permet de conserver les informations sur la longueur des arêtes (équivalentes à la valeur de BLAST).

**Détection des incongruences** Une fois que les familles d'HNP ont été détectées, il faut reconstruire chaque arbre et le comparer avec l'arbre de référence afin de détecter les incongruences. Ces incompatibilités avec la phylogénie des espèces sont causées par le fait qu'un xénologue est plus similaire au gène de l'espèce donneuse qu'aux orthologues des espèces proches de l'espèce receveuse. Ainsi les espèces impliquées dans un HGT paraîtront plus proches qu'elles ne le sont en réalité.

Les méthodes phylogénétiques constituent donc un moyen direct et naturel pour détecter les HGT. Elles reposent sur des concepts limpides et bien établis.

### 1.4.2 Méthodes non-phylogénétiques

Les méthodes de détection non-phylogénétiques peuvent être très utiles dans un contexte où les données génomiques sont limitées, soit en terme d'échantillonnage taxonomique, soit au niveau du nombre de gènes séquencés pour une espèce donnée. Cependant, avec plus de 460 génomes bactériens et archaéens séquencés en avril 2007, cet avantage devient beaucoup plus restreint. Elles permettent néanmoins d'éviter la reconstruction phylogénétique, qui peut être très coûteuse en terme de

temps de calcul surtout si l'on étudie un grand nombre de gènes et/ou d'espèces. Elles peuvent se baser sur l'analyse de la composition atypique de gènes, sur le contenu génomique et sur la discordance "phylogénétique".

### Composition atypique

Le taux de G+C est très variable chez les bactéries, allant de 25% G+C chez *Mycoplasma* à 75% G+C chez *Micrococcus* (Lawrence et Ochman, 1997). Bien que la composition en bases soit relativement homogène à l'échelle du génome, les première, deuxième et troisième positions de codons ont des compositions caractéristiques (Muto et Osawa, 1987). De plus, le contenu en nucléotides est en rapport avec la phylogénie, c'est-à-dire que le taux de G+C d'espèces proches aura tendance à être similaire, indiquant que la composition en bases peut rester stable sur de longues périodes (Ochman et Lawrence, 1996). Par conséquent, les régions dont la composition en bases ou l'utilisation des codons sont atypiques ont vraisemblablement été acquis par HGT (Lawrence et Ochman, 1997). Lawrence et Ochman (1997) détectent donc les HGT grâce à ces caractéristiques : selon leur modèle, les gènes transférés présentent un important biais de composition et un faible CAI (*Codon Adaptation Index*, voir Sharp et Li (1987)). Ils estiment à 17% la fraction d'ADN codant introduit par HGT chez *E. coli*. Ils soulignent cependant que c'est une estimation plancher car d'une part l'ADN acquis il y a suffisamment longtemps a évolué et a acquis les caractéristiques du génome hôte (processus qu'ils nomment *amélioration*), et d'autre part l'ADN acquis d'organismes ayant une composition en base similaire à *E. coli* ne sera pas détecté par leur méthode.

Hayes et Borodovsky (1998) utilisent les modèles de Markov pour modéliser les caractéristiques de composition des génomes. Par la suite, ces modèles sont incorporés dans un algorithme Bayésien qui identifie les gènes atypiques qui sont susceptibles d'avoir été acquis par HGT. Ils identifient 683 gènes atypiques chez 10 organismes, dont 176 sont significativement similaires (par BLASTP) à d'autres protéines dans la base du NCBI. Cependant le but principal de leur méthode est d'identifier des ORF (*open reading frames)* dans un génome brut.

Nakamura et al. (2004) utilisent une méthode bayésienne : après avoir modélisé les caractéristiques des régions codantes et non condantes du génome grâce aux chaines de Markov, ils expriment la probabilité postérieure qu'un fragment de nucléotide appartienne à une région codante. Ils peuvent également discriminer les gènes acquis par HGT. Leur méthode permet en outre de prédire l'espèce donneuse du gène transféré. La proportion de gènes transférés varie de 0,5% à 25%, résultats qui concordent avec d'autres études. Cependant, cette méthode présente les mêmes limitations que les autres méthodes basées sur la composition en bases, soit la non-détection de gènes acquis d'espèces avec des compositions génomiques similaires, ainsi que les transferts très anciens.

### Contenu génomique

Clarke et al. (2002) ont conçu une méthode qui détecte la discordance phylogénétique entre les gènes. Ils utilisent la méthode brh pour détecter les "orthologues" (les HNP en fait) au sein des génomes. Si tous les gènes d'un génome A ont la même histoire phylogénétique, les brh de chaque gène de A vers les autres génomes devraient avoir le même classement de brh avec les autres gènes. Si un gène présente un classement incongruent avec celui des autres gènes de A, il est phylogénétiquement discordant. En effet, par analogie avec la construction d'un arbre avec une matrice de distance, un classement de similarité contradictoire définit un arbre contradictoire (*"a conflicting pattern of similarity relationships must specify a conflicting tree"*) (Clarke et al., 2002). Ce classement incongruent (une trop grande similarité) serait causée par un HGT. Ils estiment ainsi la proportion de gènes acquis par HGT de 6,0% chez *Pyrococcus horikoshi* à 16,8% chez *Treponema pallidum*. Ces ordres de grandeur correspondent aux estimations d'autres études (par exemple Lawrence et Ochman (1997)). Cependant, la méthode est tributaire de l'identification des orthologues que l'utilisation de BLAST ne peut garantir.

### Distribution de gènes

La principale explication de la distribution observée des gènes orthologues parmi les espèces est l'héritage vertical (ou Darwinien). On s'attendrait à ce que des espèces proches partagent plus d'orthologues que des espèces plus éloignées. Cependant, l'hérédité ne prédit pas de façon parcimonieuse la distribution de certains gènes. Ragan et Charlebois (2002) réalisent une approche BLAST à partir de 23 génomes. Ils comptabilisent le nombre de hits de chaque génome en fonction des grands phyla bactériens (*Proteobacteria*, *Cyanobacteriales*, *Firmicutes*, etc.) ainsi qu'avec les autres domaines (eucaryotes et archae), et le comparent avec la distribution de hits attendue. Cette distribution est fonction du nombre de cibles potentielles (les séquences protéiques) dans les autres phyla. Ils estiment la proportion de gènes acquis par HGT de 1,9% chez *Mycoplasma genitalium* à 8,0% chez *Escherichia coli*. Comme le soulignent les auteurs, ces chiffres sont très conservateurs car ils n'incluent que les transferts inter-phylum (impliquant deux phyla) au sein des bactéries, ainsi que les transferts inter-domaine.

### 1.4.3 Limitations des méthodes non phylogénétiques

Les méthodes non-phylogénétiques présentent des avantages importants sur les méthodes phylogénétiques grâce à la rapidité de calcul ainsi que sur la quantité moindre de données nécessaires. Cependant, elles sont sujettes à plusieurs problèmes.

En premier lieu, le processus d'amélioration décrit par Lawrence et Ochman (1997) affecte les gènes acquis par transfert : leur séquence adopte la composition en bases et l'usage des codons du génome-hôte, de sorte qu'après un certain temps, ils ne sont plus discernables des gènes originaux (par le critère compositionnel). D'après les auteurs, l'amélioration devrait être plus marquée pour des gènes ayant peu ou pas de contraintes fonctionnelles. Ce phénomène affecte l'efficacité de toutes les méthodes compositionnelles, dont celles décrites précédemment.

Deuxièmement, Daubin et Perrière (2003) ont analysé la structure de l'utilisa-

tion des codons le long de plusieurs génomes bactériens : dans la plupart des phyla, ils sont très structurés avec une tendance à un enrichissement en A+T vers la région où se situe la fin de la réplication (*replication terminus*). Ceci est peut-être le résultat de contraintes liées au chromosome circulaire, et suggère que les gènes ont un contenu en bases différent en fonction de leur localisation sur le chromosome. Par conséquent, cela peut conduire les méthodes compositionnelles à surestimer les HGT. Enfin, les gènes fortement exprimés ont un biais de composition afin de maintenir une efficacité élevée lors de la traduction (Sharp et Li, 1987).

Il est donc évident qu'à l'heure actuelle, les méthodes compositionnelles sont affectées par de nombreux artéfacts qui entrainent à la fois des faux négatifs (la non-détection de gènes tranférés puis améliorés) et des faux positifs (des gènes indigènes à composition atypique). Les différentes méthodes non-phylogénétiques détectant des sous-ensembles disjoints des gènes transférés, Lawrence et Ochman (2002) recommandent l'utilisation de plusieurs méthodes pour une meilleure détection.

### 1.4.4 Comparaison des méthodes de détection

Les multiples méthodes de détection des HGT (phylogénétiques et non-phylogénétiques) ont des caractéristiques qui leurs confèrent certaines forces et faiblesses théoriques. Il est essentiel de réaliser des tests de comparaison afin de les évaluer. Cortez et al. (2005) ont mis à l'épreuve plusieurs méthodes de détection : la méthode GC de Lawrence et Ochman (1997), le profil de distribution (DP) (Daubin et al., 2003a), un modèle bayésien (BM) (Nakamura et al., 2004) et un modèle de Markov de premier ordre qu'ils ont développé (MM). Ils ont simulé le transfert de 100 gènes à partir de 30 espèces bactériennes (dont 14 protéobactéries) vers *E. coli* K12 MG1655. L'efficacité des méthodes est variable, et l'espèce donneuse a parfois une grande influence. Par exemple, pour *P. putida*, BM détecte seulement 1 à 8% des GTH (gènes transférés horizontalement) (niveaux de signification 1% et 5%), alors que MM détecte 96 à 100% des transferts. Globalement, MM détecte le plus grand nombre de transferts artificiels, avec le plus faible taux de faux positifs

(<3%). Par la suite, les auteurs ont utilisé la méthode la plus performante (MM) pour détecter les GTH réels chez deux souches d'*E. coli* et *S. typhimurium* : la plupart ne sont pas assignés à une catégorie fonctionnelle et sont principalement hypothétiques. D'autre part ils détectent des GTH appartenant à toutes les catégories fonctionnelles majeures, y compris quelques gènes informationnels comme une protéine ribosomique. Cependant la plupart semblent être des pseudogènes, que Liu et al. (2004) ont décrits comme étant principalement le résultat de HGT avortés.

Une étude similaire a été réalisée par Ragan et al. (2006)[3] qui ont comparé plusieurs méthodes paramétriques (*surrogate*) et phylogénétique. Les méthodes paramétriques sont les suivantes : la méthode GC de Lawrence et Ochman (1997) déjà testé par Cortez et al. (2005), la méthode de Hayes et Borodovsky (1998) basée sur les modèles de Markov (MM), la méthode de discordance phylogénétique (PD) de Clarke et al. (2002), la méthode de profil de distribution (DP) de Ragan et Charlebois (2002). La méthode phylogénétique est celle de Beiko et al. (2005). Les auteurs examinent la profondeur des transferts détectés (grâce à la comparaison des topologies de Beiko et al. (2005)). GC et MM identifient préférentiellement les transferts qui ont eu lieu après la divergence des entérobactéries. Ces méthodes, qui sont basées sur les caractéristiques nucléotidiques, perdent leur pouvoir de résolution à mesure que les séquences transférées sont améliorées (voir Lawrence et Ochman (1997)). PD et DP identifient plutôt les transferts d'avant la divergence des protéobactéries, et ont tendance à négliger les transferts entre espèces proches. Ragan et al. (2006) et Ragan (2001) concluent en affirmant que les méthodes non-phylogénétiques semblent détecter des ensembles disjoints de transferts, et que seules les méthodes phylogénétiques permettront à terme d'identifier les transferts de n'importe quelle ancienneté.

[3]Ragan (2001) avait réalisé une étude comparative très semblable.

### 1.4.5 Supériorité des méthodes phylogénétiques

La comparaison d'arbres phylogénétiques est en principe l'approche la plus rigoureuse pour identifier les HGT (Syvanen, 1994). En théorie, elle détecte aussi bien les transferts récents que les anciens. La seule limitation est l'impossibilité de détecter les transferts entre groupes frères, car la phylogénie n'est pas modifiée dans ce cas. Bien qu'elles puissent nécessiter beaucoup plus de puissance de calcul que les méthodes non-phylogénétiques, les méthodes phylogénétiques dépendent moins des subtilités de composition au niveau des nucléotides. De plus, les meilleures méthodes pour l'alignement des séquences et l'inférence phylogénétique sont fondées sur des modèles avec une base statistique et biologique solide, donc les cas de HGT peuvent être étudiés dans le bon contexte statistique, tant pour les hypothèses évolutives que pour les supports statistiques obtenus (Ragan et al., 2006).

La principale force des méthodes phylogénétiques est leur capacité à détecter de nombreux transferts avec une grande confiance, y compris les transferts très anciens dont les produits sont si répandus (par exemple les mitochondries) qu'on n'imaginerait pas qu'ils sont d'origine exogène (Lawrence et Ochman, 2002). Cependant, ce problème est causé principalement par les limitations des bases de données (Lawrence et Ochman, 2002), et à mesure qu'elles s'enrichissent en nouvelles espèces, la finesse de leur pouvoir de détection s'amplifie. En avril 2007, plus de 460 génomes complets de bactéries et archae étaient publiquement disponibles sur le site du NCBI. Cette masse de données, combinée à la puissance de calcul actuellement disponible, fait que les méthodes phylogénétiques deviennent de plus en plus la méthode de choix pour détecter les HGT.

## 1.5 Phylogénie et HGT

### 1.5.1 Hypothèses expliquant la fréquence des HGT

Jusqu'à présent, nous n'avons fait que constater les mécanismes et l'étendue des HGT. Nous allons maintenant voir quelles sont les hypothèses expliquant leur fréquence et leur répartition.

#### 1.5.1.1 Facteurs écologiques

Une explication naturelle de la distribution des HGT est la compatibilité écologique entre les espèces donneuse et receveuse. Comme nous l'avons déjà abordé plus haut, Jain et al. (2003) ont fait une telle étude : ils ont mis en évidence la corrélation entre HGT et divers paramètres écologiques. Autrement dit, les espèces ayant des caractéristiques similaires échangeront préférentiellement des gènes. Ce sont les caractéristiques internes comme la taille du génome, le contenu en G+C et l'utilisation du carbone (autotrophie ou hétérotrophie) qui ont la meilleure corrélation avec les HGT. Les facteurs externes comme l'oxygène (aérobie ou anaérobie) et la température de croissance ont une moins bonne associativité. Si les auteurs trouvent des explications à ces constatations, la théorie ne semble pas bien expliquer la répartition des HGT.

#### 1.5.1.2 Deux classes de gènes

Rivera et al. (1998) ont étudié plusieurs catégories fonctionnelles de gènes chez un méthanogène (*Methanococcus jannaschii*), une cyanobactérie (*Synechocystis 6803*), une protéobactérie (*Escherichia coli*) et un eucaryote (*Saccharomyces cerevisiae*). Après avoir détecté les gènes orthologues parmi ces génomes, ils recontruisent les arbres phylogénétiques correspondants. L'étude des diagrammes de dispersion des scores de similarité discrimine clairement deux classes de gènes. Il y a d'un côté les gènes impliqués dans le traitement de l'information (appartenant aux catégories fonctionnelles suivantes : transcription, traduction, replication, GTPases et ATPases, ARNt synthéthases), et de l'autre les gènes impliqués dans les opérations cellulaires (synthèse des acides aminés, biosynthèse des cofacteurs, enveloppe cellulaire, métabolisme énergétique, biosynthèse des acides gras et des phospholipides, biosynthèse des nucléotides, fonctions régulatrices). Les gènes appartenant à la première classe sont baptisés *informationnels*, alors que ceux de la seconde sont dits *opérationnels*. Les auteurs exposent d'une part l'origine chimérique des eucayotes : les gènes informationnels semblent avoir été transférés à partir du côté

méthanogène (archéen) de l'arbre alors que les gènes opérationnels proviennent plutôt du côté protéobactérien. Leur deuxième conclusion est que les gènes opérationnels semblent être plus facilement transférables, alors que ce n'est pas le cas pour les gènes informationnels, ce tant chez les eucaryotes que les procaryotes.

#### 1.5.1.3 L'hypothèse de la complexité

En poursuivant la réflexion de Rivera et al. (1998), Jain, Rivera et Lake (1999) ont proposé l'hypothèse de la complexité. Les gènes informationnels appartiennent généralement à des systèmes complexes et ont donc un nombre élevé de partenaires qui sont tous optimisés pour fonctionner de concert. Par exemple, l'assemblage de l'appareil de traduction implique plusieurs dizaines de gènes. Au contraire, les gènes opérationnels font partie d'ensembles produits par un petit nombre de gènes. Lors d'un transfert horizontal, un gène informationnel aura à établir avec succès une interaction avec un plus grand nombre de partenaires qu'un gène opérationnel. Si la probabilité $p$ d'établir une interaction avec un partenaire est $1/m$, alors la probabilité $P$ d'établir une interaction avec *tous* les nouveaux partenaires est $1/m^n$, où $n$ est le nombre de partenaires. On voit que $P \rightarrow 0$ rapidement. Ceci expliquerait pourquoi les gènes informationnels sont si peu sujets aux HGT.

Aris-Brosou (2005) a étendu hypothèse de la complexité. Il propose que les gènes dont les produits sont impliqués dans des fonctions complexes (gènes informationnels) sont moins sujets à l'adaptation évolutive. Il montre que les protéines les plus conservées ont une haute connectivité (c'est-à-dire qui ont de nombreuses interactions avec d'autres protéines), sont principalement des composants intracellulaires impliqués dans des processus et fonctions informationnels. Par ailleurs, les gènes ayant au moins un site sous évolution adaptative ont une connectivité significativement inférieure, et ont une localisation principalement membranaire ou extracellulaire. Ces résultats illustrent donc l'hypothèse de la complexité en expliquant plus en détail et plus formellement ses rouages.

### 1.5.2 Notre hypothèse : les gènes répandus sont moins sujets au HGT

Il existe un autre facteur pouvant affecter négativement le succès d'un transfert : il s'agit de la présence d'un gène ayant une fonction similaire ou identique dans le génome receveur, généralement l'orthologue du gène transféré. Pour que ce dernier puisse être fixé dans le génôme hôte, il faudrait qu'il apporte un avantage sélectif, par exemple en étant plus "efficace" que le gène résidant et ainsi le supplantant, ou bien en évoluant et en acquérant une nouvelle fonction. Cependant, ces deux scénarios sont peu probables. En effet, le gène transféré sera selon toute vraisemblance moins adapté à ses partenaires génomiques et à son environnement métabolique que le gène résidant (Jain et al., 1999), et ne sera donc pas plus performant. Notons cependant le cas des protéines ribosomiques résistantes aux antibiotiques (Brochier et al., 2002), qui apportent un avantage sélectif sans apporter de nouvelle fonction. D'autre part, comme les procaryotes ont tendance à purger le matériel génétique superflu (Kurland, 2005), le gène transféré ne sera pas conservé suffisamment longtemps pour muter et acquérir une nouvelle fonction. Le remplacement orthologue est donc un évènement peu probable. Cette constatation nous conduit à formuler l'hypothèse suivante : les gènes dont la distribution taxonomique est grande (gènes "universels" ou "répandus") devraient être comparativement moins transférés que les gènes "rares", présents chez relativement peu d'espèces. Nous nous proposons de tester cette hypothèse en sélectionnant dans un premier temps des gènes de distributions rares à universelles (des gènes présents chez un nombre restreint d'espèces aux gènes universels, en incluant ceux à distribution intermédiaire). Dans un deuxième temps, nous comptabiliserons les HGT dans chaque catégorie au moyen d'une méthode phylogénétique.

### 1.5.3 Justification de nos choix méthodologiques

Dans la section 1.4, nous avons exposé la supériorité des méthodes phylogénétiques, qui sont le moyen naturel de détection des HGT. Nous choisissons de mettre en oeuvre une telle méthode dans notre protocole. Leurs limitations sont générale-

ment mieux connues et plus étudiées que celles des méthodes non-phylogénétiques car il s'agit des limitations de l'approche phylogénétique en général, pas seulement pour détecter les transferts. C'est un domaine de recherche actif depuis plus de 30 ans (voir Felsenstein (1978)). Le principal problème auquel nous aurons à faire face est l'inconsistence de la reconstruction phylogénétique. Celle-ci est causée par le signal dit non phylogénétique (Philippe et al., 2005) : la vitesse d'évolution variable parmi les espèces (qui cause l'attraction des longues branches (Felsenstein, 2004)), l'hétérogénéité de composition en nucléotides/acides aminés (Lockhart et al., 1994) et l'hétérotachie (c'est-à-dire le changement de vitesse d'évolution à certaines positions) (Lopez et al., 2002).

Le choix d'une méthode de maximum de vraisemblance (ML) ou bayésienne permettent de minimiser l'impact de ces artéfacts (Felsenstein, 2004; Philippe et al., 2005). Cependant, leur temps de calcul peut devenir prohibitif lors du traitement d'un grand volume de données. Nous allons donc essayer de maximiser le signal phylogénétique par rapport au signal non-phylogénétique (Philippe et al., 2005) afin que des méthodes moins robustes mais plus rapides (comme les méthodes de distance) n'éprouvent pas de problèmes à inférer la bonne phylogénie. Nous choisissons des groupes d'espèces séparés par des branches relativement longues (afin d'avoir suffisamment de signal phylogénétique pour les différencier). Nous évitons les espèces avec des branches excessivement longues, mais nous sélectionnons celles ayant un ancêtre commun suffisamment ancien (c'est-à-dire que leurs branches ne sont pas trop courtes). Le nombre d'espèces par groupe sera égal afin d'avoir un arbre équilibré. Enfin, comme les relations entre les grands groupes bactériens sont très anciennes et difficiles à évaluer avec certitude (voir Brochier et al. (2002); Woese (1987)), nous choisissons de ne pas nous en préoccuper.

# CHAPITRE 2

# MATÉRIELS ET MÉTHODES

Notes :

- les programmes et scripts que nous avons créés sont identifiés par une `fonte de machine à écrire`. On pourra se reporter à la figure 2.10 pour un diagramme représentant leurs relations.
- les programmes écrits par d'autres sont dénotés par des PETITES MAJUSCULES.

## 2.1 Protocole

### 2.1.1 Données génomiques

Les génomes complets des bactéries et des archées ont été téléchargés sur le site du NCBI à l'adresse suivante : ftp ://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/ avec le script `getGenomes`. Les fichiers téléchargés sont de type faa (séquence en acides aminés), fna (séquence en nucléotides), rpt (informations sur les séquences). S'il y a plusieurs fichiers faa ou fna pour un même organisme (dans le cas où il y a des plasmides ou plus d'un chromosome), ils sont concaténés au moyen du script (`mergeGenomes`). Début avril 2007, nous avions 460 génomes complets.

### 2.1.2 Choix des espèces

Nous sélectionnons les espèces dans les grands groupes bactériens de manière à ce que (1) les branches à la base des groupes soient longues, et (2) les branches entre les espèces soient, dans la mesure du possible, un peu moins longues et équilibrées[1] (voir figure 2.1). Le premier critère est primordial, car il faut qu'il y ait assez de signal phylogénétique accumulé dans cette branche basale pour que (i) des petits gènes (~100 aa) soient capables de retrouver solidement cette monophylie, et (ii) le signal non-phylogénétique (du biais de composition, des différentes vitesses évolutives, etc.) soient toujours négligeable par rapport au signal phylogénétique.

[1] De longueur comparable.

On assemble ainsi un jeu de données de $g$ groupes avec $e$ espèces par groupe, pour un total de $g \times e = n$ espèces. Typiquement, on a de 5 à 12 groupes, et de 2 à 7 espèces par groupe.

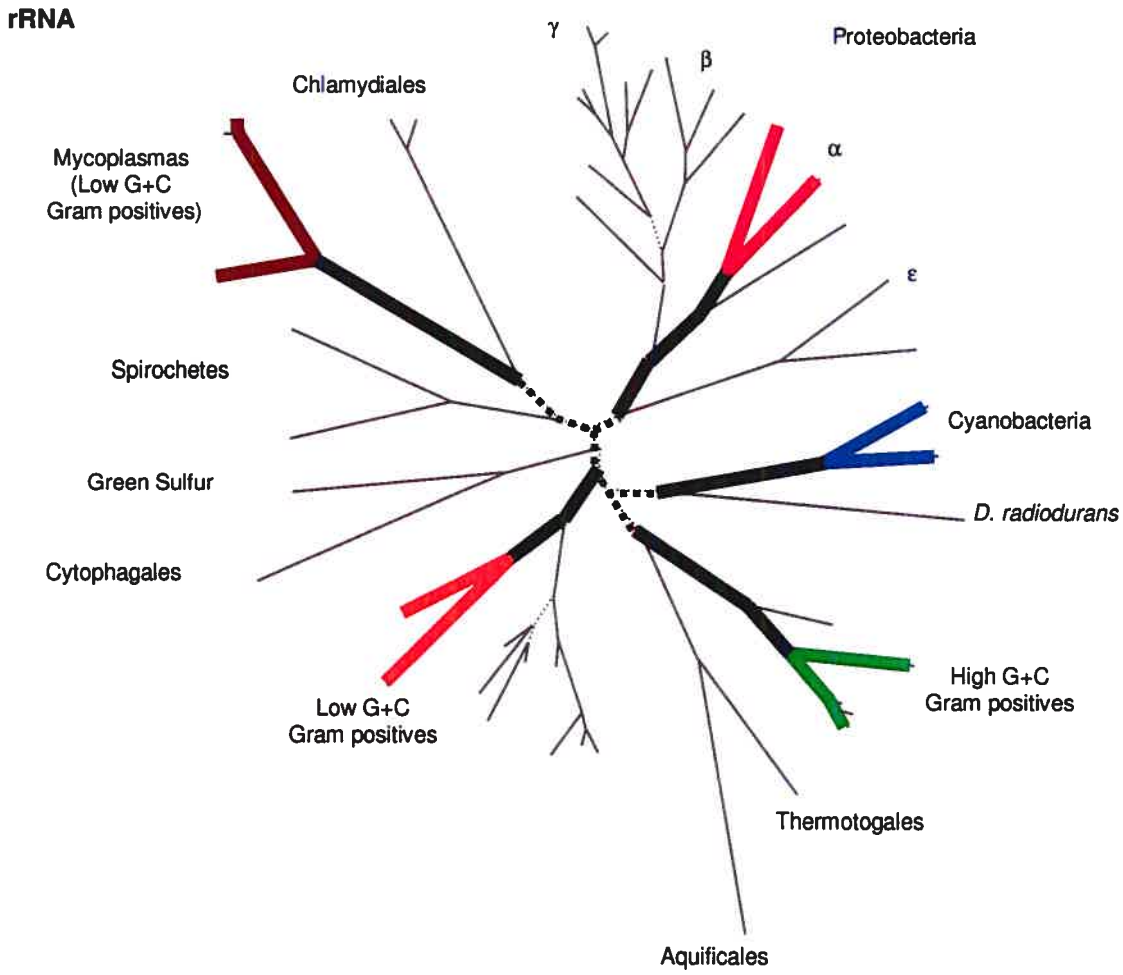


Figure 2.1 – Choix des espèces : longues branches intra-groupes (en noir), branches inter-espèces longues (en couleur). Les branches inter-groupes (en pointillé) ne sont pas considérées. Adapté de Brochier et al. (2002).

### 2.1.3 Détection des familles de gènes homologues non-paralogues

Nous avons choisi l'approche de détection par *best reciprocal hits* (brh) (Mushegian et Koonin, 1996) pour déterminer les gènes homologues non-paralogues (HNP)

parmi les espèces. Comme ces gènes sont censés se ressembler entre eux plus qu'aux autres gènes (du fait de leur origine commune), on prend les gènes les plus similaires dans chaque génome. Voici comment ils sont déterminés : à partir d'un gène $\alpha$ dans le génome A, l'outil BLAST (Altschul et al., 1990) permet de détecter les gènes les plus similaires dans un génome donné (des *hits*), qu'il classe par *e-value*[2]. Le meilleur hit[3] dans le génome B ($\beta$) a de bonnes chances d'être l'homologue du gène de départ. Cependant ce n'est pas toujours le cas (voir Koski et Golding (2001)), donc la procédure inverse est appliquée : on s'assure que le meilleur hit de $\beta$ dans le génome A est bien $\alpha$. Si c'est bien le cas, $\alpha$ et $\beta$ sont les meilleurs hits réciproques l'un de l'autre, d'où le nom de la méthode. Le principe est appliqué à tous les gènes putativement homologues dans tous les génomes du jeu de données afin de renforcer la certitude quant à l'homologie des gènes. Nous verrons plus loin comment nous nous assurons qu'ils sont non-paralogues.

Nous avons choisi d'utiliser les séquences nucléotidiques conceptuellement traduites en protéines afin de réduire certains artéfacts compositionnels. En effet la troisième position des codons est beaucoup plus variable que les deux premières en raison de la redondance du code génétique. De plus, la diversité des nucléotides est plus grande que celle des acides aminés, ce qui peut conduire à la détection de plus de dissimiliarité entre des séquences qu'il y en a réellement (Jeffroy et al., 2006). Nous avons donc utilisé la variante BLASTP de BLAST (Altschul et al., 1997).

Le processus de blast est relativement lent pour plusieurs raisons. Il nécessite beaucoup d'accès au disque, ce qui est très lent comparativement aux calculs se faisant uniquement en mémoire. D'autre part, le nombre de blasts à réaliser est élevé, car chaque génome doit être blasté contre tous les autres. Cela signifie $n \times (n-1)$ blasts de génomes entiers. Or les blasts se font gène par gène contre les autres génomes. Chaque génome contenant en moyenne 3145 gènes, on comprend aisément que l'étape du blast prenne beaucoup de temps. Pour contourner ce problème, nous avons constitué une grande base de données contenant le résultat des blasts de tous

[2]Probabilité que la similarité entre les deux séquences soit due au hasard.
[3]Celui avec la e-value la plus faible.

les génomes entre eux. Plus précisément, la base est composée de fichiers contenant le résultat du blast d'un génome contre un autre. De cette façon, lorsque nous voulons analyser un jeu de données, nous n'utilisons que les fichiers nécessaires. Nous maintenons cette base à jour à mesure que de nouveaux génomes sont publiés, ou que des génomes déjà parus changent (séquence ou annotation). Cependant, la mise à jour devient de plus en plus longue à mesure que le nombre total de génomes augmente, car chaque nouvel organisme doit être blasté contre tous les anciens, et tous les anciens doivent être blastés contre le nouveau (ce qui est en $O(n^2)$). Avec 460 génomes, il faut réaliser plus de 211 000 blasts de génome à génome, ce qui correspond à $6,64 \times 10^8$ blasts individuels[4]. Le programme principal réalisant les blasts est `blast-brh` (voir figure 2.10). Les résultats sont sauvegardés dans des fichiers resblast.

Par la suite, notre programme `brh` détecte tous les *best reciprocal hits* (brh) parmi les fichiers resblast correspondant aux génomes de notre jeu de données. Ceci donne un grand graphe non-orienté dont les noeuds sont les gènes et les arêtes les brh. `brh` sélectionne[5] ensuite tous les sous-ensembles de gènes tels que chaque gène est le brh de tous les autres gènes (voir figure 2.2). Autrement dit, `brh` détecte

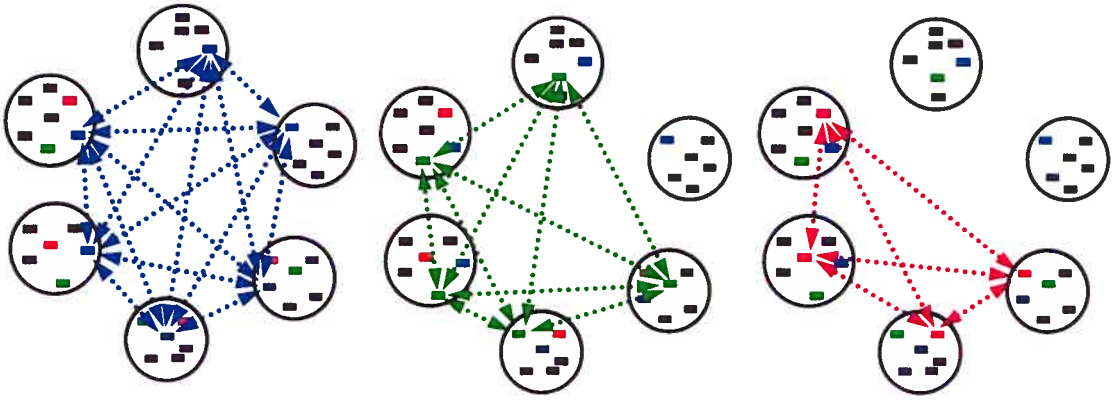

Figure 2.2 – Détection des familles de gènes homologues non-paralogues par notre programme `brh`. Dans cet exemple, il y a 6 organismes ; `brh` détecte des familles de taille 6, 5, 4.

[4] $460 \times 459 \times 3145 \approx 6,64 \times 10^8$

[5] `brh` utilise la librairie LEDA pour le traitement des graphes.

tous les sous-graphes complets[6] dans le graphe global. Les sous-graphes doivent avoir au minimum 4 noeuds[7] et au maximum $n$ (le nombre total d'espèces). Le

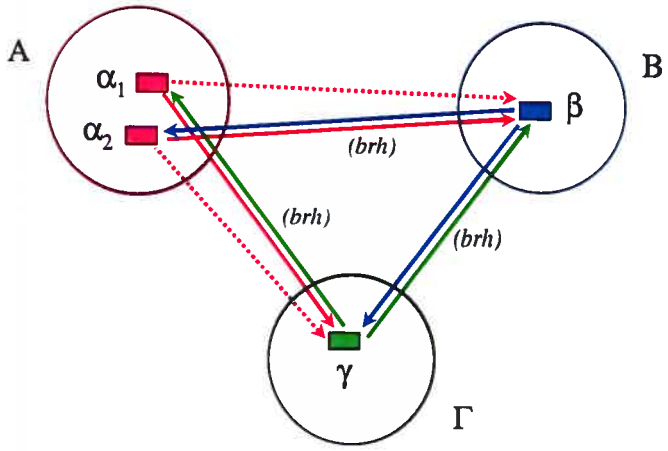


Figure 2.3 – Rejet des familles contenant des paralogues. Les arêtes représentent les meilleurs hits, pleines pour celles qui ont un meilleur hit réciproque, en pointillés pour les autres. $\alpha_1$ et $\alpha_2$ "divisent" les meilleurs hits de $\beta$ et $\gamma$. Le graphe incomplet résultant contient 4 noeuds ($\alpha_1, \alpha_2, \beta et \gamma$ ) et 3 arêtes *(brh)*. La famille est donc rejetée.

critère de complétude a pour effet d'éliminer les familles d'homologues contenant des paralogues (voir figure 2.3). Soit un génome A avec deux paralogues $\alpha_1$ et $\alpha_2$, et soient les génomes B et $\Gamma$ avec un seul homologue chacun, respectivement $\beta$ et $\gamma$. Les meilleurs hits de ces gènes vont être "divisés" entre $\alpha_1$ et $\alpha_2$ car ils sont très similaires ($\beta$ et $\gamma$ n'ayant évidemment qu'un seul meilleur hit dans A). Le graphe résultant est incomplet car il contient 4 noeuds pour seulement 3 arêtes. La famille correspondante n'est donc pas sélectionnée.

Ainsi, `brh` détecte les familles de gènes homologues non-paralogues, qui sont souvent abusivement appelés "orthologues" dans la littérature. Cette détection est rapide par rapport aux blasts : environ deux minutes pour un jeu de données avec 10 espèces, environ deux heures pour 35 espèces. Cette relative rapidité nous permet de tester facilement plusieurs combinaisons d'espèces.

[6] ou presque : nous autorisons l'absence d'un petit nombre d'arêtes ( typiquement 2) afin d'assouplir le critère de complétude. Voir sections 2.2.4 et 3.9.

[7] le nombre minimum d'espèces pour avoir un arbre phylogénétique

### 2.1.4 Alignement des séquences et sélection des régions bien alignées

Les séquences des familles d'HNP sont alignées avec MUSCLE version 3.6 (Edgar, 2004b). Ce programme réalise des alignements de qualité égale ou supérieure à CLUSTAL W (Thompson et al., 1994) tout en étant deux à trois ordres de grandeur plus rapide (Edgar, 2004a). Les options par défaut sont utilisées car cela donne l'alignement de meilleure qualité (Edgar, 2004a).

Les régions mal alignées et trop divergentes (c'est-à-dire pour lesquelles l'homologie est incertaine) sont éliminées avec GBLOCKS 0.91b (Castresana, 2000). Les options utilisées sont les suivantes : nombre minimum de séquences pour une position en bordure (*flank position*) : 75% du nombre de séquences ; nombre minimum de positions non-conservée contigües : 5 ; longueur minimum d'un bloc : 5 ; positions brèches (*gap positions*) autorisées : moitié.

Une vérification et une sélection manuelles des régions bien alignées est habituellement préférable afin de ne pas trop éliminer de positions, minimisant ainsi la perte de signal phylogénétique. Dans notre cas, la sélection manuelle est impraticable en raison du grand nombre de gènes et de jeux de données que nous souhaitons étudier. De plus, notre choix d'espèces fait que le signal devrait être suffisamment puissant pour ne pas être affecté.

Cette étape est réalisée par notre script `do_muscle_gblocks`.

### 2.1.5 Reconstruction phylogénétique

Nous reconstruisons la phylogénie pour chaque groupe d'HNP avec deux méthodes :

- **Maximum de vraisemblance** (ML) : TREEFINDER (Jobb et al., 2004), modèle WAG, $\Gamma$ 4 catégories. Le script est `do_treefinder`. Nous avons aussi écrit `do_treefinder_obelix`, une version adaptée pour l'exécution sur une grappe de calcul.
- **Distance** : Neighbor Joining avec NEIGHBOR (Felsenstein, 2005), matrices de distances estimées avec TREE-PUZZLE (Schmidt et al., 2002), modèle WAG,

$\Gamma$ 4 catégories. Le script est `do_neighbor`.

Dans les deux cas, 100 réplicats de bootstrap sont effectués avec SEQBOOT (Felsenstein, 2005). Le consensus (de type majoritaire étendu) est réalisé avec CONSENSE (Felsenstein, 2005). Nous montrerons que dans notre situation, les deux méthodes obtiennent des résultats équivalents, et nous préférerons la méthode de distance, beaucoup plus rapide (voir section 3.4).

### 2.1.6 Reconstruction de l'arbre des espèces

Afin d'obtenir l'arbre représentant nos jeux de données, nous utilisons l'ensemble des gènes présents chez toutes les espèces. Notre script `muscaphy` fait appel à `do_muscle_gblocks`) pour les aligner et sélectionner les régions bien alignées. Les gènes sont ensuite concaténés grâce à SCAFOS (Roure et al., 2007), puis TREEFINDER est utilisé pour la reconstruction phylogénétique, modèle WAG, $\Gamma$ 4 catégories, sans bootstrap.

## 2.2 Extraction des résultats

### 2.2.1 Effectifs des familles, des gènes et des groupes testables

Les effectifs des familles pour chaque taille[8] est informative : ils permettent de quantifier la proportion de gènes "rares" (présents chez 4 espèces) aux gènes "universels" (présents chez les $n$ espèces), ainsi que les gènes de répartition intermédiaire. Il est primordial d'avoir suffisamment de familles dans chaque catégorie pour la significativité statistique.

Le nombre de gènes est simplement le nombre de familles multiplié par la taille de la famille. Cela permet d'avoir une meilleure idée du nombre réel de gènes dans une taille de famille donnée. Par exemple, si les tailles 4 et 10 ont environ le même nombre de familles, la catégorie 10 contiendra en réalité 2,5 fois plus de gènes.

D'autre part, on s'intéresse au nombre de groupes réellement testables. Pour être

[8]Par taille, nous faisons référence au nombre d'espèces représentées. Exemple : "taille 4" réfère aux familles contenant 4 espèces.

testable, un groupe doit avoir au moins deux représentants, et il doit y avoir au moins deux espèces n'appartenant pas à ce groupe dans la famille d'HNP. Comme seuls les groupes testables sont pris en compte dans la comptabilisation des HGT (voir 2.2.2), cette statistique est importante. Parmi les familles avec de petits effectifs (taille 4, 5, etc.), il y a potentiellement une proportion non négligeable de groupes non testables. À l'opposé, les familles de taille $n$ seront toutes testables car toutes les espèces de tous les groupes sont représentées. C'est le script `count_effectifs` qui compte les effectifs[9].

### 2.2.2 Comptabilisation des HGT : congruence, incongruence et irrésolution

Maintenant que les arbres de chaque famille de HNP sont construits, nous devons les comparer avec l'arbre de référence[10] afin de déterminer si les espèces de chaque groupe se retrouvent bien ensemble, c'est-à-dire si chaque groupe est monophylétique (ou congruent ; voir figure 2.4), ou s'il y a des incongruences. Les incongruences sont interprétées comme le signe d'un (ou plusieurs) HGT, et non comme le résultat d'artéfacts de recontruction (à cause du choix d'espèces).

Nous devons donc examiner chaque groupe d'espèces individuellement. Si le groupe est testable (voir section 2.2.1), trois cas de figures sont possibles : il peut être **congruent** avec l'arbre de référence, c'est-à-dire que tous les représentants sont groupés ensemble avec une valeur de bootstrap supérieure à un seuil prédéterminé. Le groupe peut être **incongruent**, c'est-à-dire qu'une ou plusieurs espèces d'un autre groupe se sont insérées, ou qu'un de ses représentants est avec un autre groupe. De même, les groupements incongruents doivent avoir une valeur de bootstrap supérieure au seuil prédéterminé pour être considérés. Si un groupe a une valeur de bootstrap inférieure au seuil, il est **irrésolu** (ou non-résolu, *unresolved*).

[9]Pour compter les groupes testables, il faut avoir déterminé leur statut, ce qui est réalisé à l'étape suivante, d'où la position plus tardive de `count_effectifs` dans le diagramme 2.10.

[10]L'arbre de référence est constitué par notre connaissance à priori des groupes d'espèces. Bien évidemment, l'arbre des espèces que nous construisons pour des besoins de représentation graphique a exactement la topologie attendue, sauf peut-être pour les relations inter-phylum.

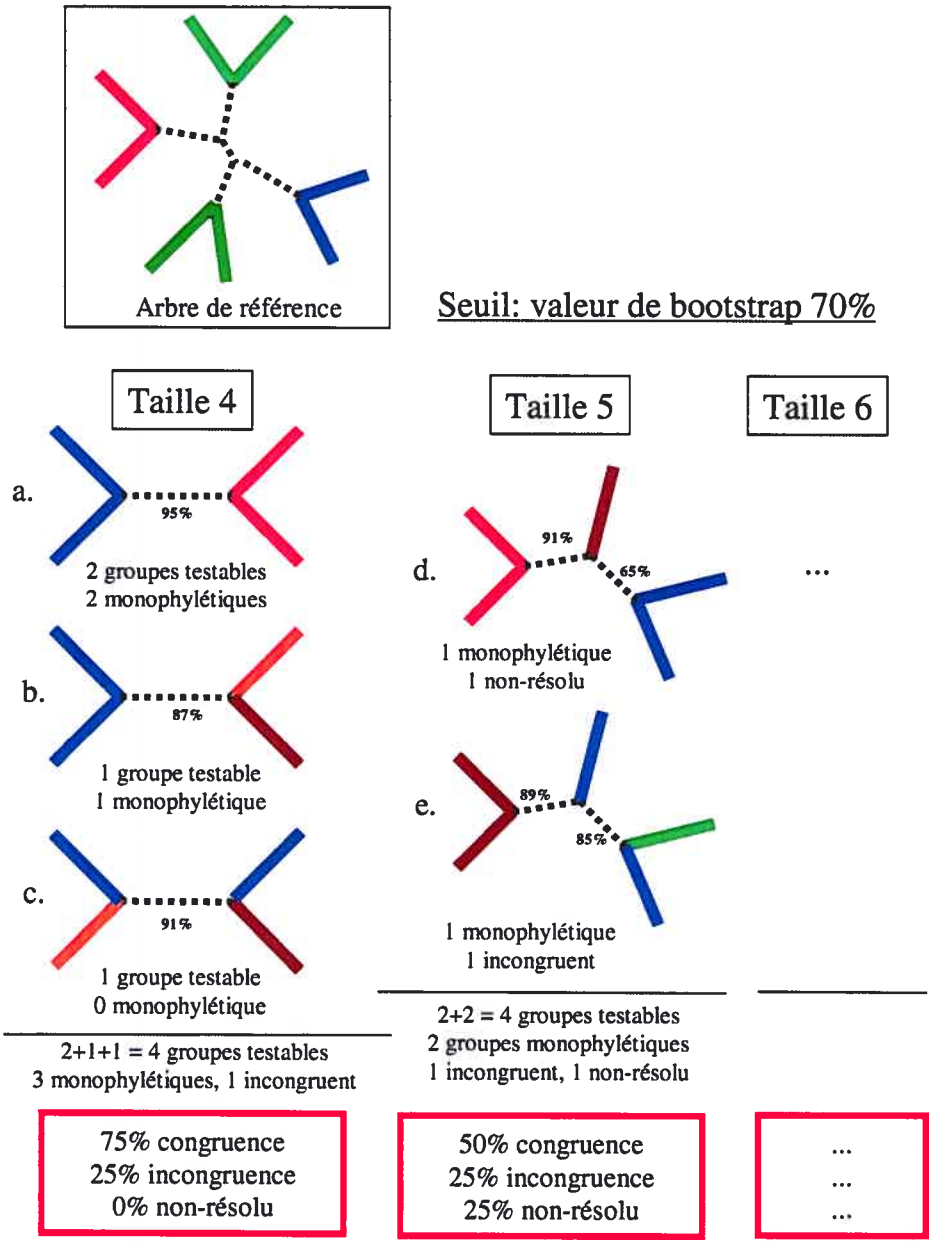


Figure 2.4 – Comptabilisation des HGT. Pour toutes les familles d'une taille donnée, on compte le nombre de groupes testables. En comparant avec l'arbre de référence, on compte le nombre de fois que ces groupes sont congruents, incongruents et non-résolus, puis on calcule le pourcentage du total.

Les groupes non-testables sont ignorés.

Par la suite, on comptabilise le nombre de fois qu'un groupe est congruent, incongruent ou irrésolu par rapport au nombre de fois que le groupe est testable, et ce pour toutes les familles d'HNP de taille 4 à $n$. On peut ensuite obtenir un pourcentage de ces valeurs pour chaque taille de famille, pourcentages qui seront représentés dans nos figures de résultats.

C'est le script `analyse_outfiles` qui réalise cette étape. Il fait appel à `interprete` qui détermine le statut de chaque famille, et `summary` qui compile les résultats.

### 2.2.3 HGT artificiels

Afin de mettre à l'épreuve la capacité de détection des HGT de notre protocole, nous avons simulé un nombre défini d'HGT dans nos données. Nous ne simulons pas l'évolution des séquences, car cela ne permet pas de bien reproduire les problèmes de reconstruction (Brinkmann et al., 2005).

Pour un jeu de données avec $n$ espèces, nous sélectionnons les familles d'HNP de taille $n$ exemptes de HGT (c'est-à-dire dont tous les groupes sont significativement congruents). La simulation d'un transfert horizontal se fait comme suit : un groupe donneur est choisi aléatoirement, au sein duquel une espèce donneuse est choisie aléatoirement (voir figure 2.5). On choisit une espèce receveuse dans un groupe receveur (différent du groupe donneur). L'espèce receveuse prend la séquence de l'espèce donneuse. Le processus est éventuellement répété, avec pour contraintes que tous les groupes donneurs sont distincts, et tous les groupes receveurs sont distincts (cependant, un groupe peut être à la fois donneur et receveur). Par la suite, chaque groupe est peuplé avec le nombre adéquat d'espèces. Évidemment, les espèces ayant "donné" leur séquence pour simuler le transfert ne sont plus utilisées. Par conséquent, les jeux de données simulés auront $g \times (e-1)$ espèces[11].

Nous effectuons également la simulation de HGT en diminuant le nombre d'espèces par groupes. Nous aurons donc des jeux de données supplémentaires avec

[11] $g$ est le nombre de groupes, $e$ le nombre d'espèces par groupe.

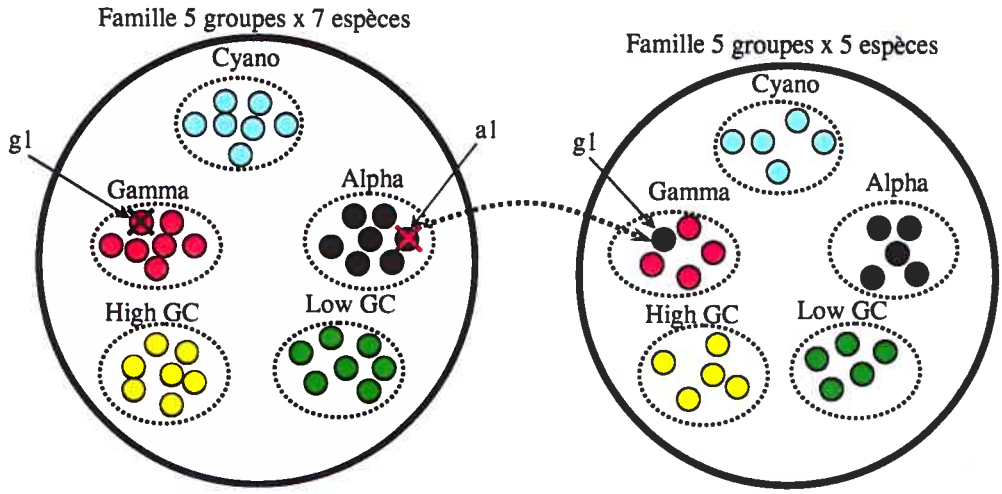


Figure 2.5 – Simulation d'un HGT à partir d'une famille 5x7 en vue d'obtenir une famille 5x5. Le groupe donneur est Alpha. La séquence donneuse est a1. Le groupe receveur est Gamma. La séquence receveuse est g1. La *séquence* de a1 prend le *nom* de g1. Une fois le transfert simulé, g1 et a1 sont supprimés de Gamma et de Alpha. Pour le peuplement des nouveaux groupes, 4 séquences de Gamma et 5 séquences de chaque autre groupe sont choisies.

$g \times (e-2)$, $g \times (e-3)$, ..., $g \times 2$ espèces. Enfin, nous répétons chacune de ces simulations 5 fois afin que des espèces différentes soient impliquées dans les transferts. Ce grand nombre de simulations nous permet d'établir des statistiques en évitant des particularités de certains gènes et/ou espèces. C'est le script `forge_brh_datasets` qui réalise les simulations en créant des nouveaux jeux de données qui sont ensuite analysés avec les outils décrits plus haut.

### 2.2.4 Seuil de e-value et arêtes manquantes dans `brh`

Le seuil d'e-value dans `brh` détermine la valeur maximale autorisée pour les deux best hits composant un brh. Nous avons testé différents seuils de e-value lors de l'exécution de `brh` : d'une valeur très laxiste (1e-1) à stricte (1e-100). Nous avons observé les effets de la variation du seuil sur le nombre de familles de gènes dans chaque catégorie. Nous avons retenu un seuil relativement laxiste de 1e-4 pour nos manipulations, car la sélection opérée par le critère de complétude de `brh` est lui strict. Afin de compenser pour la stringence de notre critère de complétude

et ainsi augmenter nos effectifs de famille de HNP, nous avons toléré un petit nombre d'arêtes manquantes dans les sous-graphes. Nous avons testé de 0 à 5 arêtes manquantes afin de déterminer la valeur optimale.

### 2.2.5 E-value des familles de HNP

La version 2.5 de notre programme `brh` détermine la e-value pour chaque famille de gènes. De tous les *best reciprocal hits* de la famille, celui dont la e-value est la plus élevée définit la e-value pour la famille au complet (voir figure 2.6). Cela nous permet de classer les familles en deux catégories en fonction de leur e-value : inférieure ou supérieure à 1e-40 [12] Nous étudions la congruence, l'incongruence et la non-résolution pour chaque catégorie. Les familles avec les meilleures e-values (inférieures au seuil) devraient présenter moins d'incongruence que celles avec les e-values plus élevées.

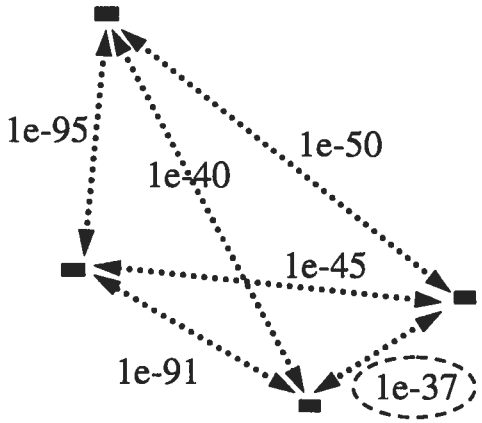


Figure 2.6 – Détermination de la e-value d'une famille d'homologues non-paralogues : l'arête (représentant un *best reciprocal hit*) avec la e-value la plus élevée (ici 1e-37) est choisie.

[12] Seuil déterminé empiriquement de manière à avoir le même nombre de familles dans chaque catégorie.

### 2.2.6 Variation du seuil de significativité du bootstrap

Comme le seuil de significativité du bootstrap est de première importance dans la détermination de nos résultats, il convient d'étudier les effets de sa variation. Nous testons des seuils clément (50%), moyen (70%) et stringent (90%).

### 2.2.7 Effet du nombre d'espèces dans les familles d'HNP

Lors de la reconstruction d'un arbre phylogénétique, le nombre d'arbres possibles augmente très rapidement. Pour 4 taxons, il existe seulement 3 arbres non-racinés. Avec 10 taxons, il y a plus de 34 millions d'arbres possibles. Ainsi pour les familles avec un nombre réduit de taxons, il y a une probabilité non-négligeable que l'on trouve le vrai arbre, purement par hasard et indépendamment des données. À mesure que la taille de la famille augmente, cette probabilité devient négligeable. Les deux protocoles suivants ("Pick from" et Rééchantillonnage) ont pour but de mieux cerner l'effet de ce biais sur nos résultats.

#### 2.2.7.1 Reconstitution de jeu de données ("Pick from")

À partir des jeux de données pour une taille particulière, on reconstitue aléatoirement des jeux de données de toutes les tailles inférieures, jusqu'à la taille minimum (4) (voir figure 2.8). Par exemple, pour les familles de taille 10, on recrée des jeux de taille 9, 8, .., 4. Puis, pour les familles de taille 9, des jeux de taille 8, 7, .., 4, et ainsi de suite. Le but de cette manipulation est de vérifier que les résultats que nous obtenons pour chaque taille de famille sont bien le résultat du signal contenu dans les données, et non l'artéfact relié à la taille de l'arbre. C'est le script `pick_from_all` qui réalise cette étape.

#### 2.2.7.2 Rééchantillonnage

Afin d'étudier le signal contenu dans chaque famille indépendamment de la taille de l'arbre, nous réalisons un tirage aléatoire de 4 gènes pour chaque famille de gènes (voir figure 2.7). On pioche donc 4 espèces pour toutes les familles de taille 5 à $n$.

On simule ainsi des familles (et donc des arbres) de taille 4 à partir des données des familles de tailles 4 à $n$. Ceci est réalisé par `interprete` qui possède des options pour choisir la taille et le nombre de répétitions du rééchantillonnage.

### 2.2.8 Catégories de longueur des gènes

La longueur des gènes devrait avoir une influence sur les résultats. En effet, les gènes plus courts contiennent comparativement moins de signal phylogénétique que les gènes plus longs. Ainsi, ils devraient être susceptibles aux erreurs stochastiques et avoir un niveau plus élevé de non-résolution. De plus, on pourra éventuellement déterminer s'ils sont plus facilement transférés (incongruence plus élevée donc) que les gènes longs. Nous séparons donc les familles en trois catégories en fonction de la longueur du gène (après alignement et sélection par GBLOCKS) de manière à ce qu'il y ait le même nombre de familles dans chaque catégorie de longueur. Nous étudions congruence, incongruence et irrésolution comme précédemment.

### 2.2.9 Raccourcissement des gènes

Dans une manipulation reliée à celles sur la longueur des gènes, nous simulons des gènes courts à partir de gènes longs. Les gènes sont classés par longueur[13] croissante et divisés en deux catégories, de 1 à $\lfloor n/2 \rfloor$ (gènes courts) et de $\lceil \frac{n}{2} \rceil + 1$ à $n$ (gènes longs)[14]. Les gènes longs sont raccourcis aléatoirement de manière à ce que leurs longueurs correspondent à celles des gènes courts. Cette étape est réalisée par le script `shorten_genes`.

### 2.2.10 Conformation des arbres et singletons

Au sein d'une taille de famille donnée, toutes les configurations de représentations taxonomiques ne sont pas équivalentes : il peut y avoir différentes combinaisons de nombre d'espèces par groupe. Par exemple, pour un jeu de données avec 2 espèces par groupe, il y a deux combinaisons possibles pour les familles de taille 4.

[13] Nombre d'acides aminés après alignement et sélection par GBLOCKS.

[14] S'il y a un nombre impair de gènes, le gène avec la longueur médiane est ignoré.

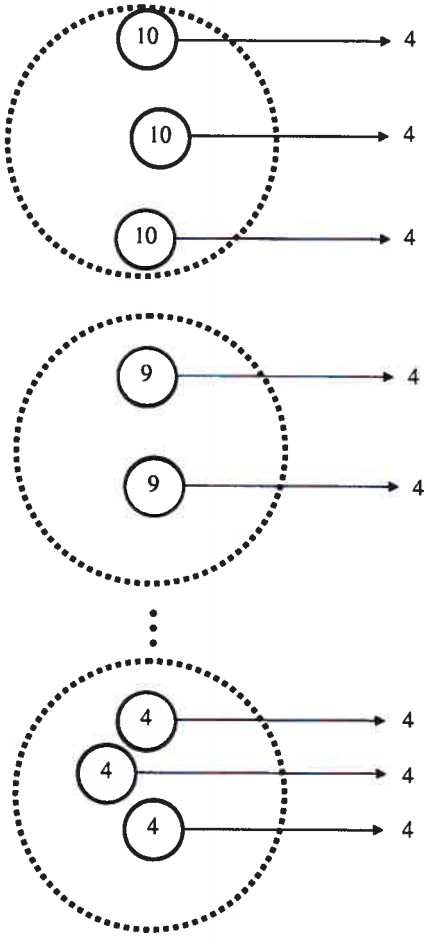


Figure 2.7 – Resampling : tirage aléatoire de 4 espèces parmi chaque famille. Les cercles en pointillés représentent l'ensemble des familles d'une taille donnée. Chaque cercle plein représente une famille de HNP contenant le nombre indiqué d'espèces.

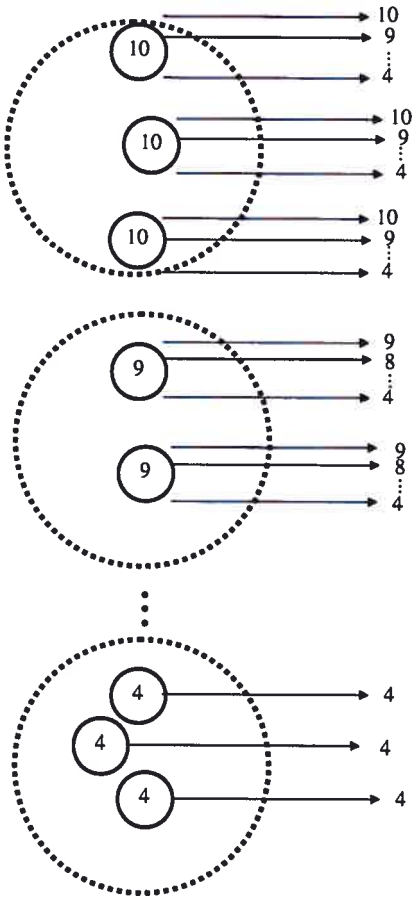


Figure 2.8 – Pick from : reconstitution de jeux de données de tailles inférieures à partir de chaque taille de famille.

Comme le montre la figure 2.9, on peut avoir 2 espèces d'un groupe et 2 espèces d'un autre (cas 2_2). L'autre possibilité est 2 espèces d'un groupe, et les deux autres de groupes différents (cas 2_1_1)[15]. Pour les familles de taille 5, les combinaisons sont : 2_2_1 et 2_1_1_1. Un gène d'une espèce qui est la seule représentante de son groupe dans une famille donnée est nommé un *singleton*. Pour les familles de taille 4, il y donc 0 ou 2 singletons. Pour les familles de taille 5, il y en a 1 ou 3. Ces singletons ont une importance particulière : en effet, comment expliquer l'absence de ce gène chez les autres espèces du groupe ? Soit le singleton a été acquis par transfert horizontal, soit l'autre espèce a perdu le gène correspondant. On s'attend donc à obtenir des résultats d'incongruence différents en fonction du nombre de singletons dans chaque famille si la présence est due à un HGT et non à une perte. Pour chaque famille, nous classons les différentes conformations en fonction de leur nombre de singletons. Nous avons développé un algorithme pour générer les différentes combinaisons en fonction du nombre de groupes, d'espèces par groupe et de la taille de la famille. Il est implémenté dans `generate_combin`, auquel notre programme `combin` fait appel afin de compter les effectifs de chaque conformation et de calculer leurs congruence, incongruence et non-résolution.

### 2.2.11 Distribution des gènes chez les espèces avec plusieurs souches

Chez certaines espèces, le contenu en gènes peut varier significativement d'une souche à l'autre (Welch et al., 2002). Certains gènes sont perdus, d'autres sont acquis par HGT ou créés par duplication. Si un gène est présent chez une ou plusieurs souches mais absent chez les autres, il est fort possible que ce gène ait été acquis très récemment par HGT. C'est ce que nous cherchons à mettre en évidence ici. À partir d'un jeu de données 5 groupes $\times$ 2 espèces, nous sélectionnons toutes les souches disponibles pour chacune des espèces. Nous partons des familles de gènes détectées par `brh`. Tous ces gènes sont blastés contre les génomes[16] des souches

[15]Il n'y a pas de cas 1_1_1_1 car il faut qu'au moins un groupe ait deux espèces afin d'être testable.

[16]Nous utilisons les séquences brutes en acide nucléiques (extension .fna). En effet, les séquences annotées (en acides aminés ou nucléiques) ont posé problème à cause d'erreurs d'annotations. Nous

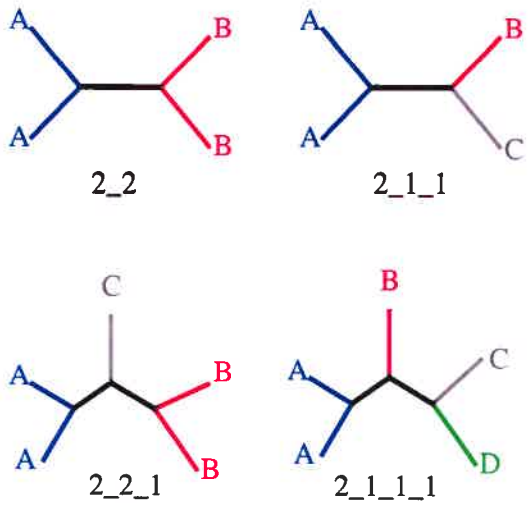


Figure 2.9 – Singletons : conformations de groupes possibles pour des arbres de taille 4 et 5, avec 2 espèces par groupe. Pour l'exemple en haut à gauche, 2_2 signifie qu'il y a 2 espèces d'un groupe et 2 d'un autre.

correspondantes afin de détecter leur présence. Un gène est dit "universel" s'il est présent chez toutes les souches d'une espèce donnée. Dans le cas contraire, il est dit "non-universel".

Les génomes de toutes les souches d'une espèce sont concaténés afin de constituer une base de données (`merge_fna`) contre laquelle `find_close_genes_N` blaste chaque famille de gènes détectée par `brh`. Par la suite, `find_blast_matches` analyse les résultats afin de déterminer si le gène est universel ou non parmi les souches (le seuil de similarité exigé pour un match est de 90%). Enfin, `correlate_incongruence` croise ces informations avec les résultats de congruence, incongruence et non-résolution déjà calculés par `analyse_outfiles`.

### 2.2.12 Retrait d'espèces

Afin de déterminer l'influence du nombre d'espèces par groupes sur nos résultats, il est nécessaire d'avoir un ensemble stable d'espèces. Nous avons utilisé notre jeu de données le plus riche en espèces par groupe, soit 5x7. Nous avons retiré une

utilisons l'option *tblastn* de BLAST puisque les gènes de départ sont en acides aminés (.faa).

espèce dans chaque groupe et ainsi obtenu un jeu 5x6. Cependant, deux options se présentent : on peut conserver les familles de HNP qui sont des sous-ensembles de celles calculées pour les 35 espèces, ou bien on peut repartir de zéro et refaire la détection des HNP avec `brh`. Nous avons réalisé les deux. Par la suite, nous avons refait l'inférence phylogénétique. Nous avons ensuite calculé congruence, incongruence et irrésolution pour ces "nouveaux" jeux d'espèces. Le processus entier est répété pour obtenir successivement des jeux 5x5, 5x4, 5x3 et 5x2.

### 2.2.13 Tirages aléatoires d'espèces

Les résultats obtenus pour un échantillon d'espèces donné peuvent être influencés par des particularités dues aux espèces choisies, par exemple la taille du génome, le taux de G+C, etc. Afin de minimiser ces effets, nous constituons un pool d'espèces à partir duquel on réalise des échantillons aléatoires d'un nombre fixe de groupes et d'espèces par groupe. Nous avons réalisé 100 tirages de 5 groupes avec 2 espèces chacun.

### 2.2.14 Diagramme des programmes

La figure 2.10 représente schématiquement le flux de données dans l'ensemble de nos programmes. Ils sont écrits en C++, bash et perl.

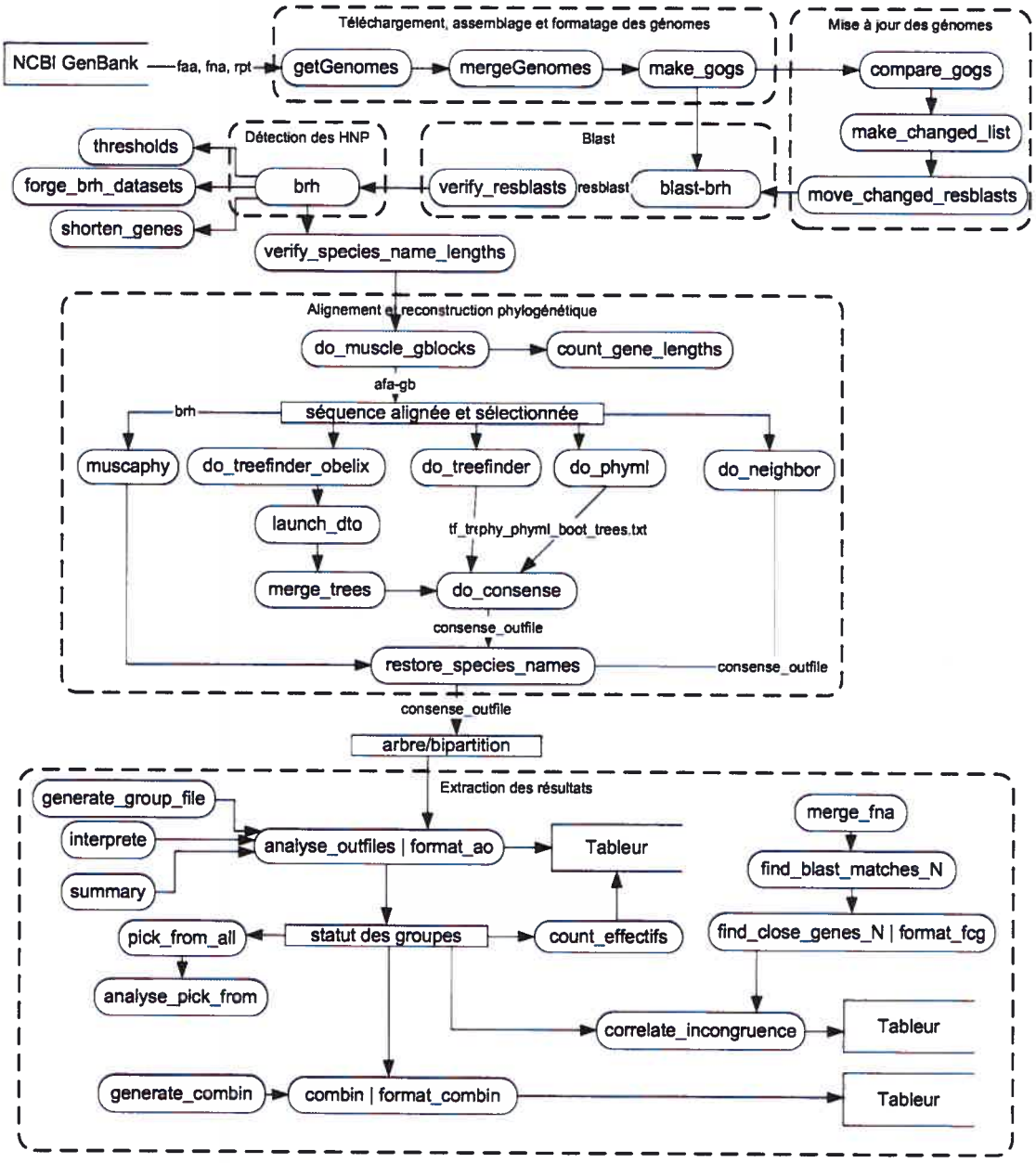


Figure 2.10 – Diagramme des programmes

# CHAPITRE 3

# RÉSULTATS ET DISCUSSION

Notes : désignation des jeux de données : [nombre de groupes]x[nombre d'espèces par groupe]. Exemple : 5x2 signifie un jeu de données formé de 5 groupes, avec 2 espèces par groupe.

Légende des figures : Nom du jeu de données, seuil e-value pour `brh`, méthode de reconstruction, seuil de bootstrap. Si non précisées, les valeurs par défaut sont **1e-4** pour le seuil d'e-value, **distance** pour la méthode de reconstruction et **70%** pour le seuil de bootstrap.

## 3.1 Exemple d'application de notre protocole : le jeu de données 5x7

Nous avons constitué un riche jeu de données de 5 groupes, avec 7 espèces par groupe (5x7), selon les critères décrits dans le chapitre Matériels et méthodes. L'arbre des espèces est montré à la figure 3.1. Le premier résultat concerne les effectifs des familles de HNP (homologues non-paralogues) (voir figure 3.2). On remarque que la majorité des familles sont de taille 4 à 7, représentant souvent des gènes spécifiques à un groupe. À l'opposé, il y a une centaine de familles de gènes présents chez toutes les espèces, des gènes universels donc. Les effectifs pour les gènes de taille intermédiaire sont très faibles. Le graphique du nombre total de séquences par famille nous montre une répartition plus équilibrée : les familles de taille 8 à 14 représentent de 132 à 248 gènes. Les autres tailles intermédiaires (15 à 34) contiennent plusieurs dizaines de gènes chacune, parfois plus de 200 (la taille 31 en contient 279 ; la taille 34, 374). La taille 35, contenant les gènes présents chez toutes les espèces, contient 3710 séquences, ce qui en fait de loin la plus grande (elle contient 75% plus de séquences que la deuxième, la taille 4, qui en contient 2100). Cette catégorie représente à elle seule près de 28% du nombre total de séquences détectées par `brh` (13270). Cela nous indique que les gènes universels à l'échelle de nos 35 espèces sont majoritaires. Enfin, le nombre de groupes testables, qui rappelons-le donne les effectifs réels sur lesquels sont exprimés les taux de congruence, d'incongruence et d'irrésolution, montre que les familles de petite

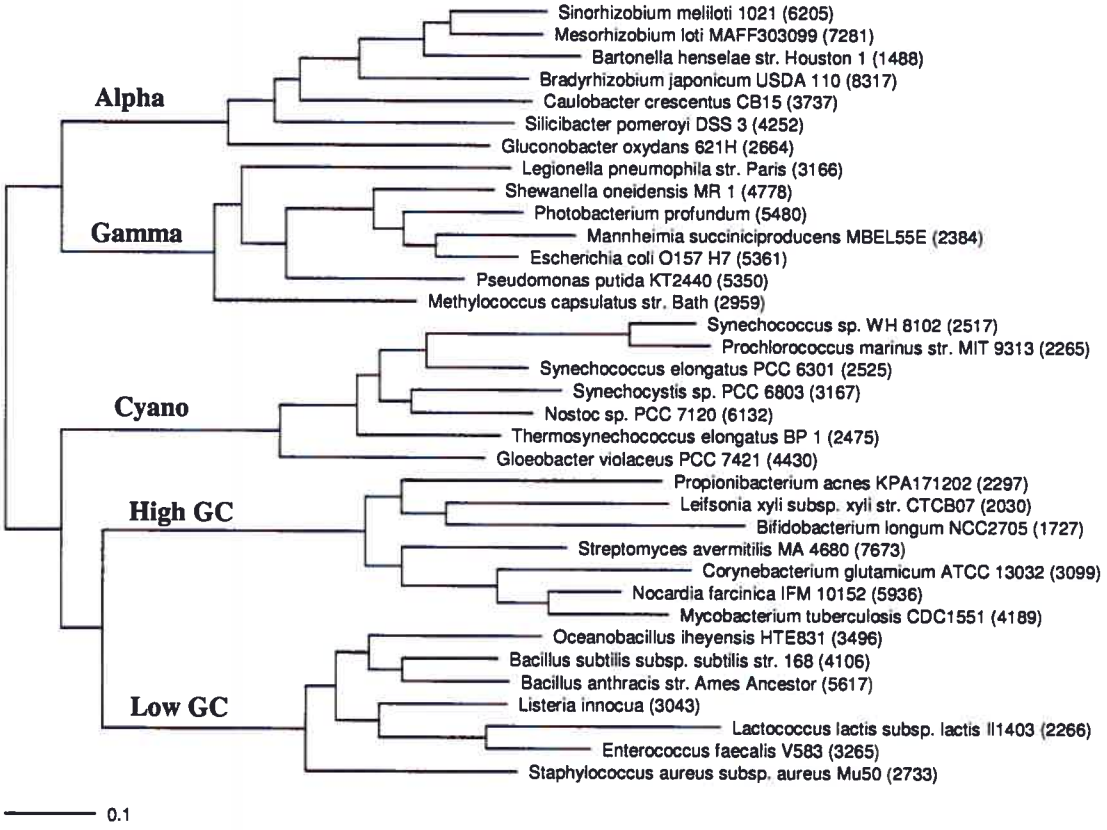


Figure 3.1 – Arbre du jeu de données 5x7. Le nombre de gènes pour chaque espèce est donné entre parenthèses.

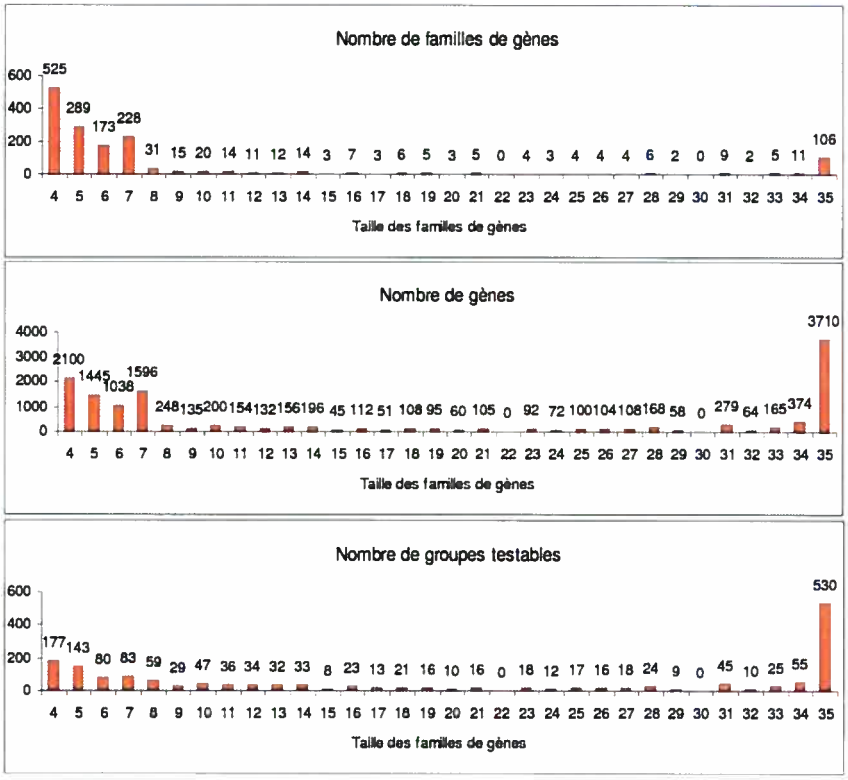


Figure 3.2 – 5x7 : effectifs des familles de gènes, nombre total de gènes et nombre de groupes testables.

taille (4-7) contiennent proportionnellement moins de groupes testables que les familles de tailles supérieures, ce qui diminue leur poids relatif. Ceci est attendu, car beaucoup de ces familles contiennent des gènes spécifiques à un groupe, ce qui les rend non-testables d'un point de vue phylogénétique[1]. À l'opposé, dans les familles de grande taille (31-35), les cinq groupes sont testables (ex : taille 31 : 9 familles, 45 groupes testables ; taille 32 : 2 familles, 10 groupes testables ; etc.). En conclusion, les nombres de groupes testables ont une distribution plus équilibrée que le nombre de familles et le nombre de gènes, ce qui est avantageux pour notre protocole.

L'analyse de la congruence, de l'incongruence et de l'irrésolution, réalisée en maximum de vraisemblance, avec WAG+Γ4 et un seuil de bootstrap de 70% (voir figure 3.3), montre une grande variabilité des taux, surtout pour les tailles de familles intermédiaires (8 à 34). Cette variation stochastique n'est pas surprenante

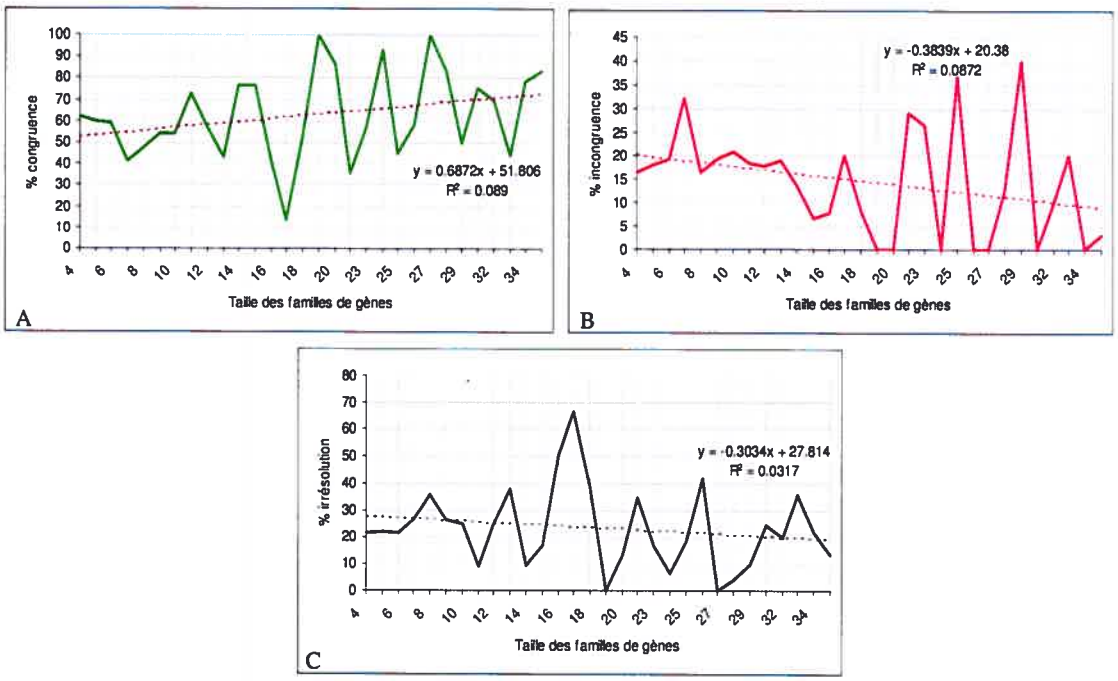


Figure 3.3 – 5x7 : congruence, incongruence et irrésolution, nombre de groupes testables. 1e-4, ML, 70%. Les points correspondant aux familles de taille 22 et 30 ne sont pas représentés car leur effectif est 0 (voir figure 3.2).

[1]Pour être testable, un groupe ne doit pas être le seul représenté dans une famille, autrement dit il doit être accompagné d'espèces d'(un) autre(s) groupe(s) (voir section 2.2.1).

vu la faiblesse des effectifs pour ces tailles. Malgré cela, on remarque que la courbe de tendance de l'incongruence (en pointillé) a une pente négative, ce qui confirme à première vue notre hypothèse. En effet, les gènes rares (tailles 4 à 7) ont de 15 à 20% d'incongruence, alors que les gènes universels ou presque (tailles 34 et 35) présentent moins de 3% d'incongruence. Autrement dit, les gènes rares présentent plus d'incongruence due aux HGT que les gènes universels. Parallèlement, la congruence augmente avec la diffusion[2] des gènes, ce qui est attendu.

Enfin, il est surprenant de constater que l'irrésolution baisse pour les familles de grande taille, alors qu'on attendait le contraire à cause du nombre plus élevé de noeuds à résoudre. Cette baisse inattendue de l'irrésolution, mise en parallèle avec la baisse de l'incongruence, pourrait également être interprétée comme étant causée par des HGT, et plus précisément des HGT provenant d'espèces non représentées dans notre jeu de données. En effet, pour que notre protocole détecte des HGT, il doit y avoir un groupement incongruent qui soit suffisamment supporté (c'est-à-dire avec une valeur de bootstrap supérieure au seuil). Ceci ne peut se produire que si le gène de l'espèce receveuse provient d'une espèce appartenant à un des autres groupes de notre jeu de données, ou du moins qu'elle en soit suffisamment proche phylogénétiquement (voir figure 3.4 A). Si ce n'est pas le cas, le gène de l'espèce receveuse n'aura d'"affinité" avec aucun groupe et donnera lieu à de l'irrésolution (figure 3.4 B). Ainsi, il est probable qu'au moins une partie de l'irrésolution que nous obtenons soit causée par des HGT de provenance extérieure à notre jeu de données, le reste étant le fruit du signal non-phylogénétique ("bruit"), ou plus simplement d'un signal phylogénétique trop faible.

## 3.2 HGT artificiels et efficacité du protocole

Afin de montrer que notre protocole détecte vraiment les HGT (c'est-à-dire qu'ils se traduisent bien par de l'incongruence), nous avons simulé des HGT ar-

[2]Par diffusion nous entendons l'étendue de la distribution d'un gène parmi les espèces, par analogie avec la diffusion d'un journal : un journal à grande diffusion est plus répandu, plus universel qu'un journal à faible diffusion.

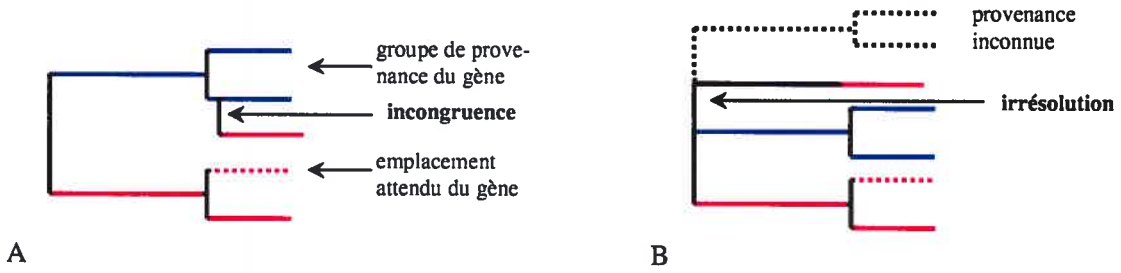


Figure 3.4 – Irrésolution causée par des HGT. A. Si le gène transféré provient d'un groupe présent dans le jeu de données, il produira un groupement significativement incongruent. B. Si le groupe donneur n'est pas représenté, le gène transféré produira un groupement irrésolu à la base de l'arbre.

tificiels dans ce jeu de données. À partir des familles de gènes présents chez les 35 espèces exempts de HGT (c'est-à-dire retrouvant les 5 groupes avec une valeur de bootstrap $> 70\%$), nous avons simulé 1, 2 puis 3 transferts (voir figure 3.5). On voit que les niveaux d'incongruence augmentent de façon marquée avec chaque

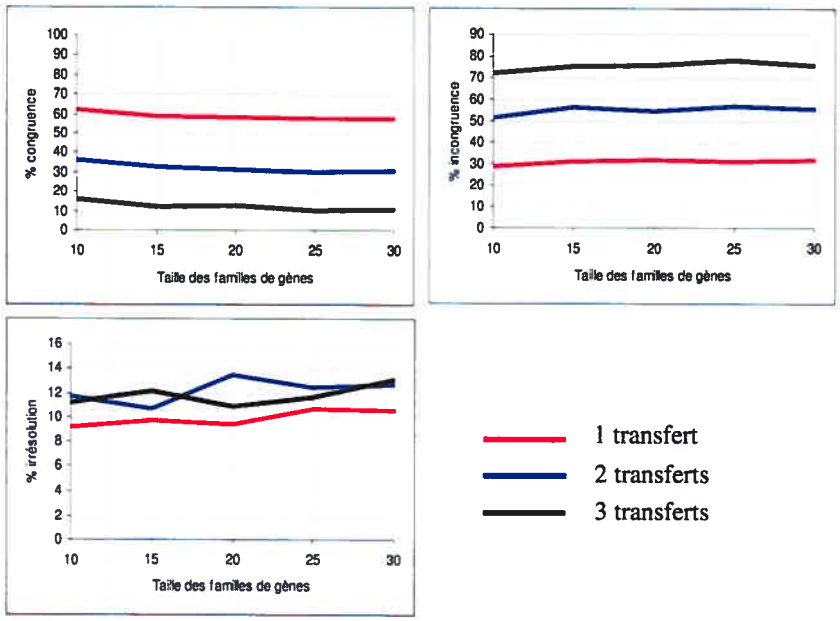


Figure 3.5 – HGT artificiels à partir d'un jeu de données 5x7.

HGT simulé, passant à 30%, 55% et 75% respectivement. Par contre, on ne peut pas directement lier ces pourcentages au nombre de groupes rendus incongruents par les simulations de transfert. En effet, avec un transfert, le groupe donneur et

le groupe receveur devraient être incongruents, donnant un pourcentage de 40%. D'autre part, notre protocole simule des transferts très récents, puisque nous n'altérons pas les séquences transférées afin de simuler les mutations et l'amélioration au nouveau génome. Néanmoins, ces résultats montrent clairement que les HGT se traduisent par de l'incongruence. L'irrésolution est presque équivalente (et faible) quel que soit le nombre de transferts simulés. Le fait que des HGT au sein même du jeu de données n'affecte pas l'irrésolution est compatible avec notre hypothèse des HGT comme cause partielle de l'irrésolution.

## 3.3 Homogénéisation de la taille des familles : jeu de données 5x2

Pour notre jeu de données 5x7, les effectifs des gènes de diffusion intermédiaire étaient trop faibles pour être statistiquement significatifs. Nous avons donc réduit le nombre d'espèces par groupe. Nous avons assemblé un jeu de données 5x2 (voir figure 3.6). Cela a aussi l'avantage de réduire considérablement le temps de calcul.

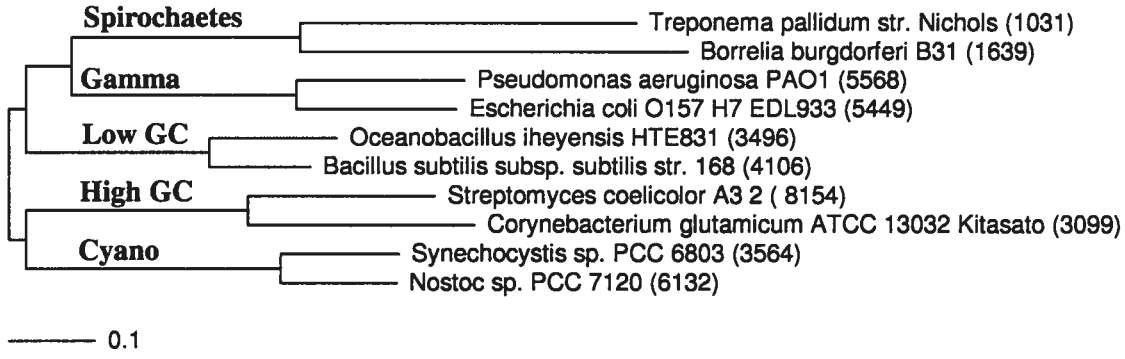


Figure 3.6 – 5x2 original : arbre.

Comme le montre la figure 3.7, les tailles avec le plus grand nombre de séquences sont encore la plus petite (4 espèces avec 900 séquences) et la plus grande (10 espèces avec 1330 séquences). Même s'ils sont plus faibles, les effectifs pour les tailles intermédiaires devraient être suffisamment nombreux pour être significatifs.

Les fluctuations des courbes de congruence, incongruence et irrésolution sont toujours présentes (voir figure 3.8). Elles sont moins prononcées que pour le jeu

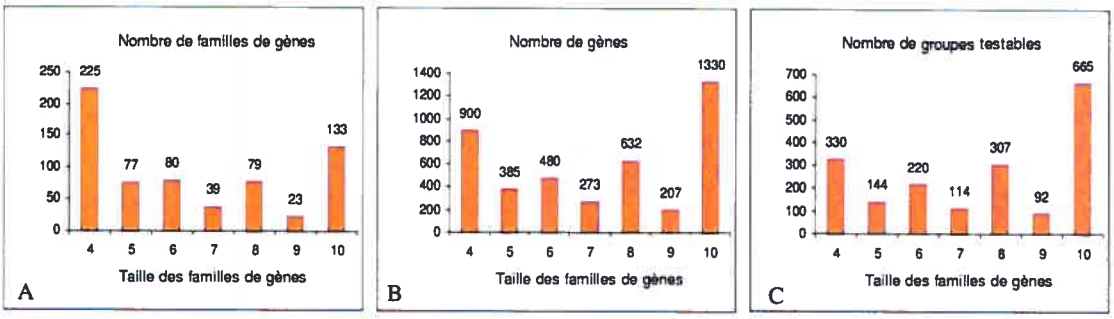


Figure 3.7 – 5x2 original : effectifs des familles de gènes, nombre total de gènes et nombre de groupes testables.

5x7, et semblent avoir une périodicité de 2. Plus précisément, il semble y avoir des pics d'incongruence pour les tailles impaires (5, 7 et 9). La courbe d'incongruence (B) présente la même pente négative et la même forme que pour le jeu 5x7. Il en va de même pour la courbe de congruence (A), dont la pente est positive comme la congruence du jeu 5x7. Par contre, la courbe d'irrésolution (C) a une pente très légèrement positive, à l'opposé du jeu 5x7, ce que nous ne comprenons pas. La pente négative de l'incongruence confirme cependant celle obtenue avec le jeu 5x7, confortant notre hypothèse.

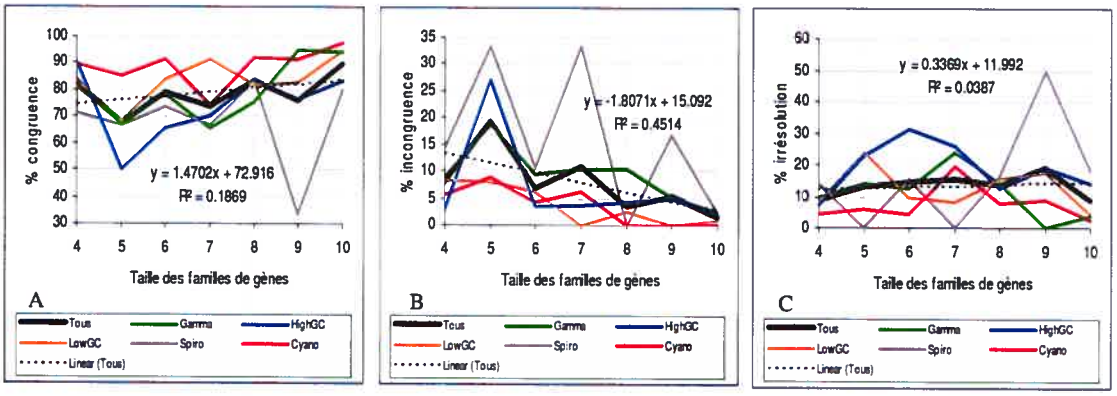


Figure 3.8 – 5x2 original, 1e-4, ML, 70%. Les courbes moyennes sont représentées en gras.

En plus des taux moyens, nous avons représenté les courbes pour chaque groupe (figure 3.8). On remarque que les spirochètes ont une incongruence au dessus de la moyenne, alors qu'à l'opposé les cyanobactéries présentent très peu d'incongruence.

Les spirochètes de ce jeu de données ont donc un taux élevé de HGT, alors que les cyanobactéries en ont peu. Il ne s'agit cependant que de deux espèces dans chaque cas, et nous ne nous prononçons pas pour l'ensemble des espèces de ces groupes.

## 3.4 Comparaison des méthodes de maximum de vraisemblance et de distance : jeu 5x2

La reconstruction phylogénétique avec la méthode de maximum de vraisemblance (ML) est la meilleure méthode d'un point de vue théorique (Felsenstein, 2004) et elle permet d'éviter plusieurs artéfacts (Jeffroy et al., 2006). Cependant, elle présente l'inconvénient d'être lente. Comme notre choix d'espèces vise à maximiser le signal phylogénétique, les méthodes moins précises mais plus rapides devraient être en mesure de retrouver la bonne phylogénie. Nous comparons ici la méthode ML utilisée jusqu'à présent, avec une méthode de distance plus rapide. Les résultats obtenus avec cette méthode sont très similaires à ceux de la méthode ML (voir figure 3.9), surtout pour l'incongruence, qui est le paramètre le plus important dans notre cas car il mesure exclusivement les HGT. Il est important de noter que l'irrésolution est sensiblement différente. En effet, la résolution est nettement supérieure avec les distances qu'avec le ML. Cela est vraisemblablement dû à une plus petite variance pour ces méthodes (Nei, 1996).

Ce résultat montre qu'il n'y a pas de problème de reconstruction et que même une méthode moins perfectionnée fonctionne bien dans notre cas, ce qui entrainera un gain de temps considérable lors de la reconstruction phylogénétique (qui est l'étape la plus longue dans notre protocole).

## 3.5 Les résultats irréguliers ne sont pas causés par la taille et la forme des arbres ("Pick from")

Les résultats en dents de scie pour la congruence, l'incongruence et l'irrésolution que nous obtenons sont peut-être causés par la taille et la forme mêmes des arbres. En effet, il est plus facile de retrouver un petit arbre car il y a moins de noeuds à ré-

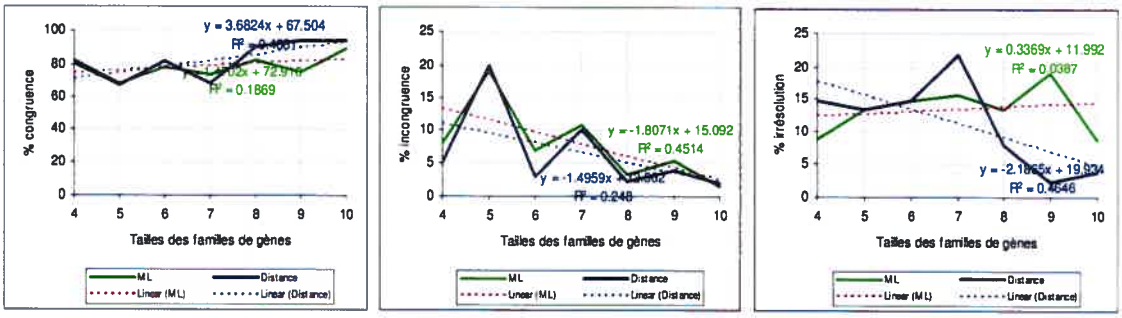


Figure 3.9 – 5x2 original, 1e-4, 70%. Comparaison des méthodes de ML et distance.

soudre que dans un arbre avec beaucoup d'espèces. Par ailleurs, l'arbre à retrouver est plus asymétrique avec un nombre impair de séquences qu'avec un nombre pair. A partir de chaque taille de famille, nous simulons des jeux de données de toutes les tailles inférieures pour tester cette éventualité. Les courbes sont remarquablement plates à l'exception du cas à 4 espèces (voir figure 3.10). Cela signifie que les résultats sont bien dus au signal contenu dans chaque arbre, et non à leur taille ou à leur forme. Par contre, on remarque une anomalie pour la taille 4 : la congruence

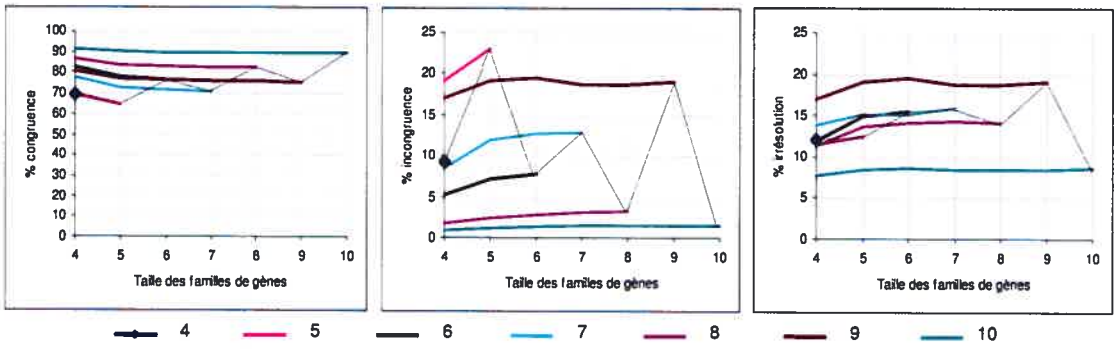


Figure 3.10 – Pick from : 5x2, 70%, 1000 réplicats. À partir des données dans chaque taille de famille, on recompose des jeux de données de tailles inférieures.

est plus élevée alors qu'incongruence et irrésolution sont plus faibles, et ce pour les données provenant de toutes les tailles. L'explication de ce phénomène se situe dans la topologie des arbres de taille 4 : ceux-ci n'ont qu'une seule branche interne, donc une seule valeur de bootstrap. Le problème survient si l'on a deux espèces appartenant à deux groupes, l'un monophylétique (la valeur de bootstrap étant

supérieure au seuil se significativité), le deuxième étant représenté par des espèces ayant des gènes homologues non-paralogues peu semblables, par exemple si une des deux espèces l'a acquis par HGT. Dans ce cas, le deuxième groupe sera considéré monophylétique "par défaut" puisque les deux conditions pour la monophylie que nous avons définies[3] sont remplies. Deux groupes réellement monophylétiques donnent un arbre avec une longue branche interne, et des branches terminales comparativement courtes (figure 3.11, à gauche). L'arbre du cas problématique aura une branche interne plus courte, avec les branches terminales relativement longues pour les espèces du groupe monophylétique "par défaut" (figure 3.11, à droite). Notre méthode est incapable de discriminer entre les deux cas, ce qui conduit à

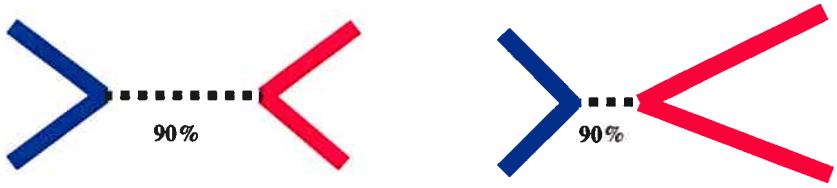


Figure 3.11 – Artéfact pour les arbres de taille 4 : à gauche, deux groupes véritablement monophylétiques ; à droite, seul le groupe bleu est monophylétique. Le groupe rouge est trouvé monophylétique, mais de façon erronée (voir texte).

une surestimation de la congruence pour les familles à 4 séquences, et donc une sous-estimation de l'incongruence et de l'irrésolution. Pour remédier à ce défaut, il faudrait examiner les longueurs de branches, mais déterminer les rapports de longueurs qui permettraient de différencier les "vraies" monophylies des "fausses" ne serait pas aisé.

Grâce à notre choix d'espèces, il y a suffisamment de signal phylogénétique pour que la bonne phylogénie soit inférée quelle que soit la méthode, quel que soit le nombre d'espèces et quelle que soit la forme de l'arbre. Par contre, l'incongruence (ou la congruence) ne s'applique pas bien au cas 4 espèces où la congruence est surestimée. Il faudra donc toujours considérer les valeurs à 4 séquences avec précaution.

[3] Au moins 2 espèces pour le groupe en question avec au moins 2 espèces d'autres groupes, et valeur de bootstrap supérieure au seuil de significativité.

## 3.6 Impact de la longueur des gènes

Les gènes ont été classés en 3 catégories afin de vérifier si les gènes courts ont moins de signal et sont plus sujets au HGT que les gènes longs. Le résultat n'est pas très concluant, comme attendu. L'irrésolution est plus élevée pour les gènes courts (voir figure 3.12, en rouge), ce qui indique un signal phylogénétique plus faible. Par contre, l'incongruence n'est pas différente pour les gènes moyens et longs. Donc la longueur d'un gène ne semble pas influencer sa propension à être transféré.

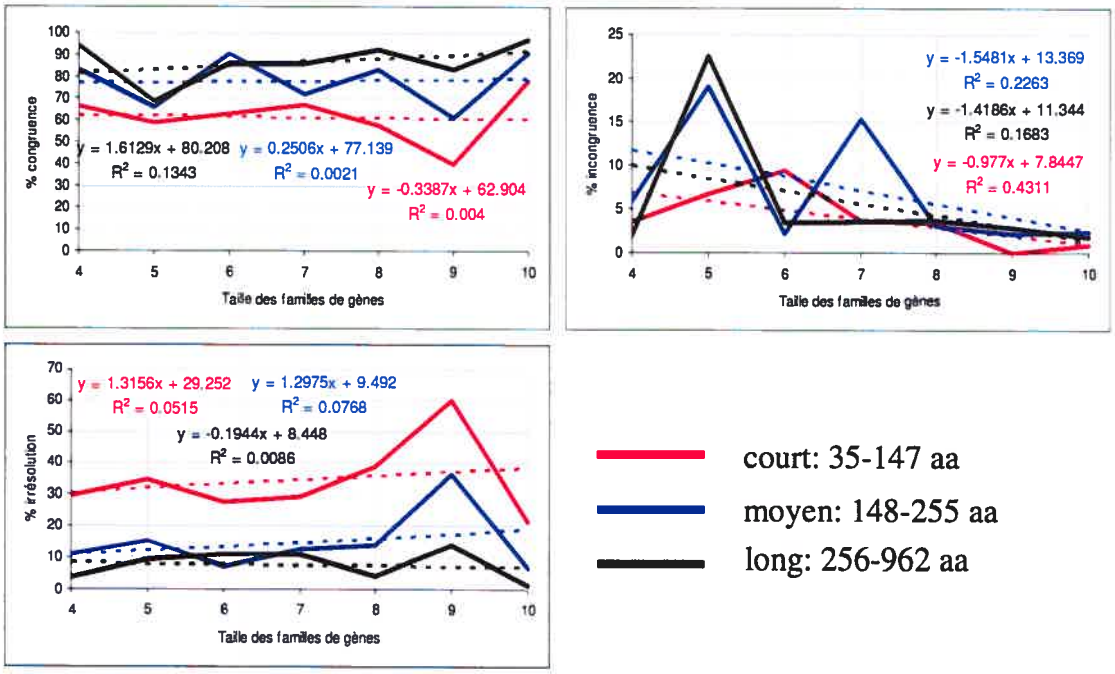


Figure 3.12 – Longueur des gènes : congruence, incongruence et irrésolution en fonction de la longueur des gènes.

La manipulation complémentaire qui consiste à simuler des gènes courts à partir de gènes longs n'est pas plus concluante. Nous n'observons pas de différence significative entre les gènes courts et les gènes longs raccourcis (voir figure 3.13). Ces résultats présentent tout de même plusieurs caractéristiques incongrues. Pour les gènes courts, la taille 4 présente des taux d'incongruence (B) et d'irrésolution (C) très faibles, alors que la congruence est élevée (85%). Une explication possible serait une plus grande proportion de familles dans lesquelles le deuxième groupe est

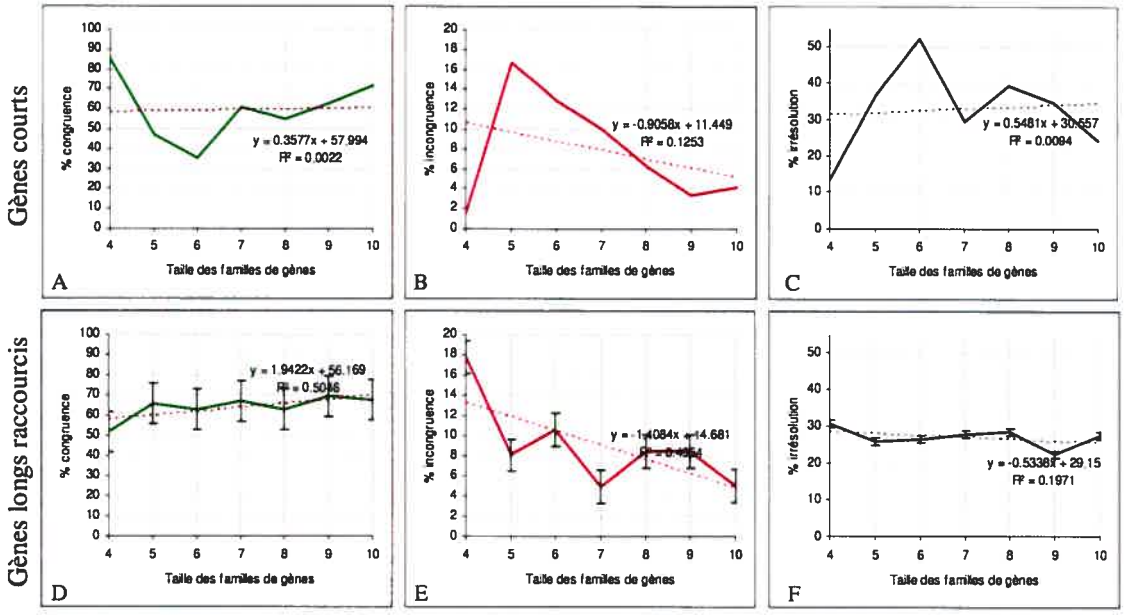


Figure 3.13 – Raccourcissement : la moitié des gènes les plus longs sont réduits aléatoirement à la taille de la moitié des gènes les plus courts (10 tirages).

monophylétique par défaut, tel que décrit dans la section 3.5. Cette interprétation est complémentée par les taux de la taille 4 pour les gènes longs raccourcis : la congruence (D) est faible (50%), et l'incongruence et l'irrésolution sont élevées. Ce type de résultat où les gènes les plus rares sont les plus incongruents est conforme à la prédiction de notre hypothèse initiale. Cependant, nous n'avons pas d'explication tangible pour la façon dont cette discrimination aurait pu s'effectuer par le seul raccourcissement des gènes, et sans apparaître dans les gènes longs non-raccourcis (figure 3.12). De plus, les pics d'incongruence sont inversés par rapport à précédemment, les tailles impaires (5 et 7 surtout) ayant une incongruence plus faible que les tailles paires (6 et 8). Finalement, le résultat le plus net est l'augmentation de l'irrésolution pour les gènes longs raccourcis (F). Si le raccourcissement supprime la même proportion de signal phylogénétique (SP) et non-phylogénétique (SNP), n'affectant donc pas le rapport SP/SNP, le montant brut de signal phylogénétique restant n'est peut-être plus assez élevé pour inférer pleinement la congurence et l'incongruence, résultant en une hausse de l'irrésolution.

## 3.7 Influence du seuil de bootstrap

Nous analysons la congruence, l'incongruence et l'irrésolution en faisant varier le seuil de significativité afin de déterminer une valeur acceptable. Évidemment, ce seuil affecte de façon directe les courbes obtenues (voir figure 3.14). Les courbes de

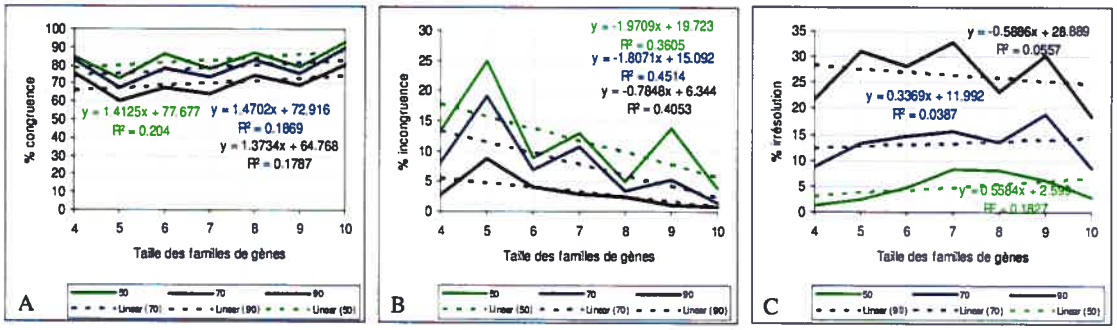


Figure 3.14 – Congruence, incongruence et irrésolution à différents seuils de bootstrap (50%, 70%, 90%), 1e-4

congruence (A) sont très ressemblantes, avec les mêmes pics pour les valeurs paires. Plus le seuil est élevé, moins les pics d'incongruence sont marqués (B), la courbe pour 90% étant presque plate. C'est l'inverse qui se produit pour l'irrésolution, où c'est la courbe pour 50% qui est dépourvue de pics. Comme nous l'avons montré dans la section 3.5, les pics d'incongruence et d'irrésolution sont bien causés par un signal contenu dans les données. Nous avons testé le seuil de 50% à des fins exploratoires seulement ; en effet, la probabilité que l'arbre inféré ne soit pas le bon est élevée (0,5). Quant au choix entre les deux autres seuils, il va influer sur notre interprétation des HGT : le seuil de 90%, privilégiant l'irrésolution, indiquerait une prédominance des HGT de provenance extérieure à notre jeu de données. Le seuil de 70% qui préserve à la fois les pics d'incongruence et d'irrésolution donnerait une image plus équilibrée. De plus, Hillis et Bull (1993) ont montré qu'une valeur de bootstrap de 70% correspond à une probabilité de 95% que le clade soit vrai. C'est donc le seuil que nous avons sélectionné.

## 3.8 Le rééchantillonnage des familles montre que les courbes sont le résultat du signal phylogénétique

Afin d'étudier le signal contenu dans chaque famille indépendamment de la taille de l'arbre, nous fixons la taille de l'arbre à 4. Cela permet de comparer le signal de chaque taille dans un contexte phylogénétique identique. En particulier, cela permet d'inclure les familles à 4 séquences qui, comme nous l'avons vu à la figure 3.10, n'ont pas les mêmes caractéristiques de détection que les autres tailles. Pour chaque famille de gènes, quelle que soit sa taille, nous tirons aléatoirement 4 espèces (ce processus est répété 1000 fois). Les courbes obtenues sont identiques à celles de notre résultat 5x2 original (voir figure 3.8), un peu plus lisses certes, mais présentant les même pics pour les tailles impaires. Cela confirme que les pentes des courbes sont bien le résultat du signal phylogénétique contenu dans les données et non d'un artéfact dans la méthode de détection.

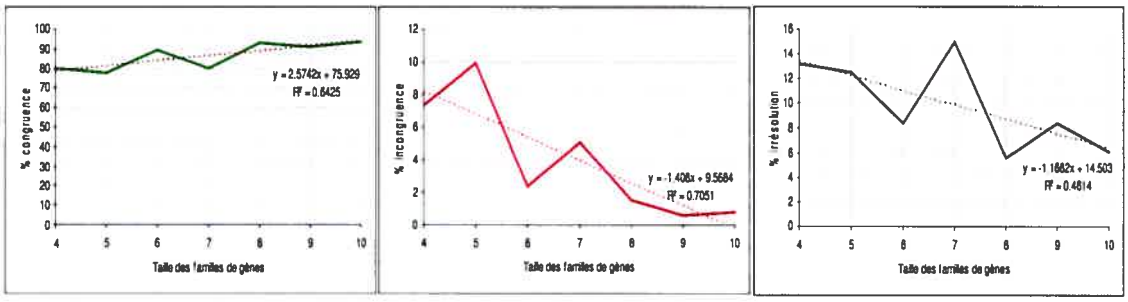


Figure 3.15 – Rééchantillonnage de 5x2 original : 4 espèces pour chaque taille, 1000 réplicats.

## 3.9 Influence du seuil d'e-value et du nombre d'arêtes manquantes tolérées sur la détection des familles de HNP par `brh`

Les résultats obtenus jusqu'à présent dépendaient des familles de HNP sélectionnées par **brh**, mais nous n'avons pas étudié comment les paramètres de ce programme influent sur les effectifs des HNP. Les deux paramètres sont le seuil d'e-value pour les *best reciprocal hits* (brh) d'une part, et le nombre d'arêtes man-

quantes (des brh) tolérées dans les sous-graphes (les familles de HNP) d'autre part. Le tableau 3.1 montre la répartition des familles en fonction du seuil d'e-value

| | log e-value | | | | | | |
|---|---|---|---|---|---|---|---|
| Taille des familles de gènes | -1 | -2 | -4 | -10 | -20 | -50 | -100 |
| 4 | 125 | 142 | 153 | 167 | 151 | 97 | 35 |
| 5 | 47 | 63 | 61 | 62 | 52 | 33 | 14 |
| 6 | 58 | 66 | 67 | 56 | 52 | 37 | 13 |
| 7 | 26 | 33 | 34 | 33 | 28 | 15 | 5 |
| 8 | 74 | 76 | 70 | 68 | 56 | 34 | 11 |
| 9 | 22 | 24 | 20 | 17 | 16 | 9 | 2 |
| 10 | 131 | 128 | 125 | 112 | 82 | 36 | 17 |
| Total | 483 | 532 | 530 | 515 | 437 | 261 | 97 |

Tableau 3.1 – Nombre de familles de gènes en fonction des seuils d'e-value pour `brh`

exigé pour les brh. On constate logiquement que le seuil le plus stringent (1e-100) donne le nombre le plus faible de familles de HNP (97). À mesure que l'on relaxe le seuil, le nombre de familles obtenues augmente jusqu'à un maximum de 532 pour un seuil de 1e-2. Si on relaxe un peu plus le seuil (1e-1), on obtient par contre moins de familles. Cela est probablement causé par le fait qu'à ce seuil, les niveaux de similarité exigés pour un brh sont trop faibles, et certains gènes détectés ne sont pas homologues. Ces faux positifs "cassent" les relations de brh dans les sous-graphes, conduisant à leur rejet par le programme `brh`. Dans la littérature, Beiko et al. (2005) ont utilisé une valeur de 1e-2 alors que Clarke et al. (2002) ont choisi 1e-10. Un seuil de 1e-4 (Cortez et al., 2005) nous semble un bon compromis en limitant les faux positifs tout en donnant un plus grand nombre de HNP. De toute façon, n'importe quelle valeur entre 1e-2 et 1e-10 donne des résultats virtuellement identiques, donc l'influence de ce choix arbitraire sur nos résultats finaux sera très faible.

Bien que le choix d'un seuil d'e-value de 1e-4 ait augmenté le nombre de HNP obtenus, les effectifs des familles de tailles intermédiaires sont encore parfois trop faibles, surtout lorsqu'il y a beaucoup d'espèces dans le jeu de données, par exemple pour le cas 5x7 (voir figure 3.2). Nous avons donc assoupli la rigueur de la détection

| Taille des familles de gènes | Nombre d'arêtes manquantes 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 4 | 153 | 181 | 223 | 229 | 229 | 229 |
| 5 | 61 | 71 | 78 | 87 | 91 | 92 |
| 6 | 67 | 74 | 79 | 80 | 88 | 90 |
| 7 | 34 | 39 | 39 | 42 | 43 | 47 |
| 8 | 70 | 75 | 79 | 80 | 80 | 83 |
| 9 | 20 | 22 | 23 | 26 | 26 | 26 |
| 10 | 125 | 129 | 133 | 138 | 140 | 140 |
| Total | 530 | 591 | 654 | 682 | 697 | 707 |

Tableau 3.2 – Nombre de familles de gènes en fonction des arêtes manquantes tolérées pour **brh** à un seuil d'e-value de 1e-4.

effectuée par **brh** en autorisant l'absence d'un certain nombre d'arêtes (brh) dans les familles de HNP. Nous avons choisi d'en tolérer 2, car comme nous allons le voir, cette valeur procure un bon compromis. Le tableau 3.2 montre les effectifs obtenus en fonction du nombre d'arêtes manquantes. Tolérer une arête manquante entraîne une augmentation d'environ 10% du nombre de familles (530 à 591). En tolérer 2 apporte une nouvelle hausse de 10% (591 à 654). Par contre, tolérer une troisième n'apporte que 4% de familles en plus. On voit que par la suite les effectifs n'augmentent presque plus à mesure que l'on tolère plus d'arêtes manquantes. En tolérer 2 s'avère donc un compromis raisonnable. Tous les résultats obtenus dans ce travail (y compris ceux présentés auparavant) ont été obtenus avec ce paramètre. Évidemment, ce paramètre peut sembler un peu simpliste. Le fait de pouvoir spécifier un pourcentage d'arêtes manquantes qui dépendrait du nombre total d'arêtes dans le graphe des brh serait préférable à un nombre brut, mais n'apporterait pas d'amélioration notable : en effet, le nombre de familles avec beaucoup d'arêtes manquantes est faible : par exemple, il y a seulement 6 familles de taille 4 avec 3 arêtes manquantes, et seulement 2 familles de taille 10 avec 4 arêtes manquantes.

## 3.10 Classification des familles de HNP en fonction de leur e-value

Nous nous intéressons à présent à l'e-value des familles de HNP *après* leur détection par `brh`.

| Taille | <1e-40 | >1e-40 |
|---|---|---|
| 4 | 332 | 168 |
| 5 | 382 | 207 |
| 6 | 363 | 162 |
| 7 | 365 | 227 |
| 8 | 303 | 156 |
| 9 | 351 | 224 |
| 10 | 322 | 167 |
| Moyenne | 345 | 187 |

Tableau 3.3 – Longueurs moyennes des familles (acides aminés après sélection par GBLOCKS) avec e-value inférieure à 1e-40 et supérieure à 1e-40.

L'e-value d'une famille est la e-value maximale parmi ses brh (voir figure 2.6). Les familles avec une e-value proche de 0 sont composées de gènes longs et/ou très conservés, c'est-à-dire beaucoup de positions et/ou pas de signal phylogénétique, ce qui est contradictoire. Les familles avec une e-value proche de 1 contiennent des gènes courts et/ou à évolution très rapide, ce qui implique peu de signal phylogénétique et/ou beaucoup de signal phylogénétique (mais potentiellement trop à cause de la saturation). Le tableau 3.3 nous montre qu'effectivement les familles avec une e-value inférieure à 1e-40 sont près de deux fois plus longs (345 positions) que les familles dont l'e-value est supérieure à 1e-40 (187 positions).

On peut émettre l'hypothèse que les gènes dans les familles avec une e-value plus élevée, étant plus courts et évoluant plus rapidement, seraient impliqués plus fréquemment dans des HGT que les gènes plus longs et conservés appartenant aux familles dont la e-value est faible. En effet, leur longueur réduite pourrait faciliter leur transfert d'un point de vue physique (via un phage ou directement dans l'environnement par exemple), et d'autre part leur taux d'évolution rapide favoriserait leur adaptation au génome hôte, et donc leur fixation.

Les résultats que nous obtenons (voir 3.16) ne vérifient pas cette hypothèse

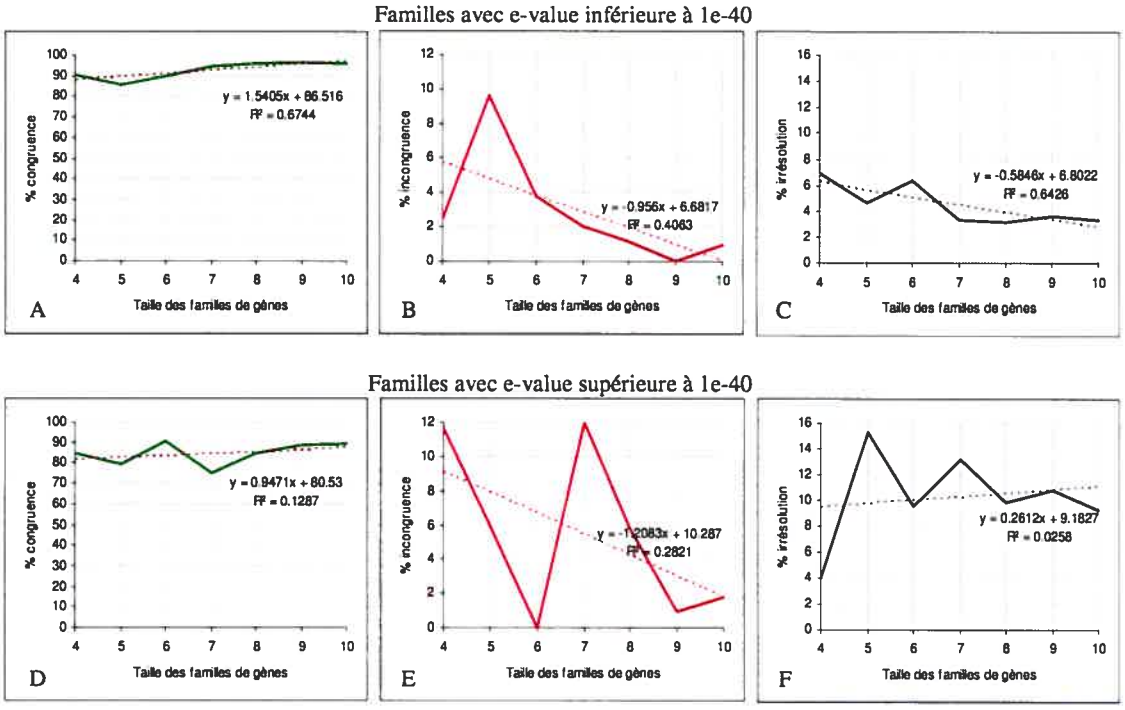


Figure 3.16 – E-value des familles de HNP : les familles sont réparties en 2 catégories en fonction de leur e-value. Les familles avec les meilleures e-values (inférieure à 1e-40) sont en haut ; celles avec les moins bonnes en bas (supérieure à 1e-40).

de façon claire. La congruence pour les familles avec les meilleures e-values (A) est un peu plus élevée, et leur incongruence (B) plus faible que ceux des familles avec une e-value élevée (D et E). Par contre, ces dernières ont une irrésolution (F) sensiblement plus élevée. La saturation élevée ou la taille réduite des gènes de ces familles semblent donc influencer à la hausse leur irrésolution. Cependant, il faudrait réaliser des manipulations supplémentaires pour tester cela.

Cette expérience nous montre que les différences constatées en fonction des e-values sont faibles, et donc que le choix arbitraire du seuil de 1e-4 pour les brh n'a que peu d'influence sur nos résultats.

## 3.11 Le cas des singletons

Certaines familles de HNP contiennent une seule espèce d'un groupe. Pourquoi les autres espèces de ce groupe ne sont-elles pas représentées ? Il y a deux explications : soit le gène de l'espèce seule (un *singleton*) a été acquis par HGT, soit les autres espèces du groupe ont perdu leur exemplaire du gène. Si le gène a été perdu chez certaines espèces, il n'empêchera pas celles l'ayant conservé d'être monophylétiques. Par contre, si le gène a été acquis par HGT, il a de bonnes chances d'être groupé avec des séquences appartenant à celui de la source. Il pourra aussi donner lieu à de l'irrésolution si le groupe donneur n'est pas représenté.

Afin de tester cela, nous avons classé les familles en fonction de leur conformation, c'est-à-dire le nombre d'espèces de chaque groupe présentes dans la famille. Par exemple, pour une famille de taille 4, et avec 2 espèces par groupe, les deux conformations possibles sont 2_2 (deux espèces dans deux groupes) et 2_1_1 (deux espèces d'un groupe, et deux autres de deux groupes différents). Ici nous regroupons les familles en fonction du nombre de singletons qu'elles possèdent. Dans notre exemple, les deux conformations 2_2 et 2_1_1 ont respectivement 0 et 2 singletons.

Le premier résultat est constitué par les effectifs de chaque conformation. Alors que les figures d'effectifs précédentes nous avaient laissé l'impression d'une certaine uniformité au sein des tailles de familles (voir par exemple la figure 3.7), la figure

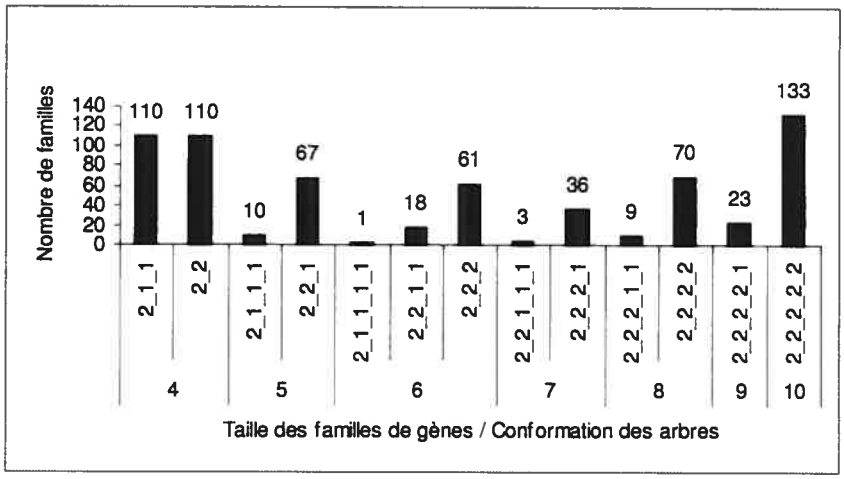


Figure 3.17 – Singletons 5x2 original : effectifs

3.17 nous montre au contraire la variété des conformations dans chaque taille. Parmi les familles de taille paire, les familles sans singletons sont plus nombreuses que celles avec deux singletons. Par exemple, pour la taille 8, il y a 70 familles 2_2_2_2 contre seulement 9 du type 2_2_2_1_1. Une exception frappante est pour la taille 4, au sein de laquelle il y a un nombre égal des deux conformations 2_2 et 2_1_1 (110). Ce résultat est plutôt contre-intuitif, car on s'attendrait à ce que les familles 2_1_1 soient moins fréquentes : en effet, les HGT ou les pertes qui les engendrent devraient être plus rares que les gènes présents chez deux espèces par groupes (2_2). Cela pourrait être dû à une particularité du jeu de données, mais d'autres jeux montrent la même tendance, le nombre de 2_1_1 dépassant même parfois celui de 2_2 (résultats non montrés). Ces effectifs doivent varier en fonction de la e-value et du nombre d'arêtes manquantes, mais nous n'avons pas mené ces tests complémentaires.

Examinons congruence, incongruence et irrésolution non pas en fonction des tailles de familles comme nous les avons présentées jusqu'à présent, mais en fonction du nombre de singletons dans chaque famille (voir figure 3.18). Les familles sans singleton (2_2, 2_2_2, 2_2_2_2, 2_2_2_2_2) présentent un taux élevé de congruence (88%,) mais surtout un taux d'incongruence remarquablement faible de moins de 4%. Les familles avec un singleton présentent un profil assez semblable, l'incon-

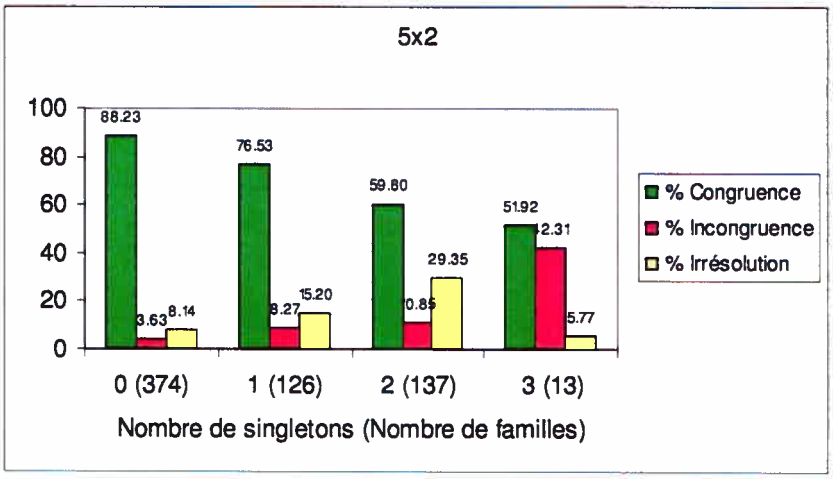


Figure 3.18 – Singletons 5x2 original : congruence, incongruence et irrésolution en fonction du nombre de singletons par famille.

gruence étant tout de même doublée (8%). Ce résultat est compatible avec la dualité de l'hyphothèse : le singleton peut être autant dû à une acquisition par transfert dans l'espèce seule (dans quel cas le gène aura de grandes chances de causer de l'incongruence) qu'à une perte dans l'espèce soeur, ce qui ne causera pas d'incongruence. Pour les familles à 2, et surtout à 3 singletons, la probabilité d'une perte concourante d'un même gène dans différents groupes est plus faible. La distribution des gènes à travers les groupes seulement chez certaines espèces (autrement dit la présence de plusieurs singletons) devient plus parcimonieusement expliquée par des HGT. Ceci se reflète dans nos résultats puisque les familles avec 3 singletons présentent un taux d'incongruence de 42% associé à une faible irrésolution (6%). Ce signal détecté par notre protocole indique clairement des HGT. De plus, on remarque que la congruence diminue régulièrement, ce qui est logique si l'on admet que l'irrésolution peut être causée par les HGT. Il aurait fallu faire la manipulation suivante : créer des familles à 2 ou 3 singletons à partir des familles à 0 et à 1 singleton.

Le nombre de conformations possibles pour un jeu de données 5x2 est faible (13). Pour un jeu de données 5x7, il y en a 774. Nous étudions donc les singletons dans notre jeu 5x7 présenté précédemment. Le grand nombre de combinaisons

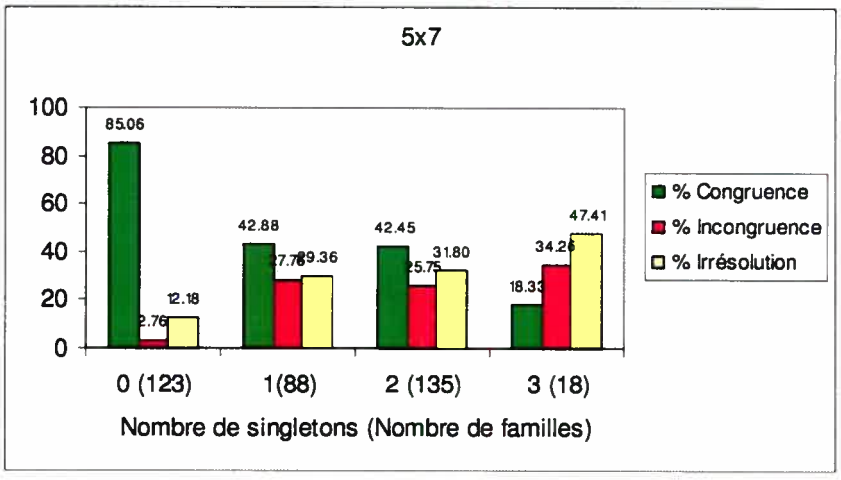


Figure 3.19 – Singletons 5x7 : congruence, incongruence et irrésolution en fonction du nombre de singletons par famille.

possibles autorise une meilleure granularité pour les conformations des familles. Évidemment, les effectifs de chaque conformation sont souvent très faibles, mais il est possible de les regrouper selon le critère du nombre de singletons, comme précédemment. Nous avons par exemple 33 conformations avec un singleton, parmi lesquelles 2_2_1, 4_3_2_2_1, 7_7_1, 7_5_4_3_1 et 7_7_7_7_1. Les résultats de congruence sont similaires à ceux obtenus avec le jeu 5x2 (voir figure 3.19). Pour les familles sans singletons, nous n'avons retenu que les familles avec des groupes complets (7_7, 7_7_7, 7_7_7_7 et 7_7_7_7_7). Les résultats de congruence, incongruence et irrésolution sont virtuellement identiques à ceux des familles sans singletons pour le jeu 5x2. Cela confirme notre supposition que les gènes universels (de l'échelle groupe-spécifique à bactérie-spécifique) sont très peu sujets aux HGT. Quant aux familles avec 1 et 2 singletons, elles présentent des taux d'incongruence nettement plus élevé que leur équivalents du jeu 5x2. L'explication probable est que beaucoup de ces conformations contiennent aussi des "doubletons", c'est-à-dire des gènes présents chez seulement deux espèces. Comme dans ce jeu nous avons 7 espèces par groupe, la perte de l'homologue chez les cinq autres espèces peut être assez improbable[4]. Comme dans le cas des singletons, la présence de plus d'un doubleton

[4]Cette probabilité est dépendante de la phylogénie à l'intérieur du groupe. Par exemple si les

renforce l'hypothèse d'un HGT.

Ainsi, l'étude des différentes conformations d'arbres nous a révélé d'importantes disparités potentielles entre des familles de taille identique, mais avec des nombres de singletons différents.

## 3.12 Absence des gènes souches proches des espèces de référence

Cette expérience reprend le même principe que la précédente avec les singletons, mais à l'échelle des souches d'une même espèce. À l'heure actuelle, il existe un bon nombre d'espèces ayant plusieurs souches dont les génomes sont entièrement séquencés. Cela nous permet de déterminer si les gènes présents chez certaines souches mais absents chez d'autres présentent plus d'incongruence que les gènes présents chez toutes les souches. Dans le cadre de cette section seulement, les premiers seront nommés "non-universels", les seconds "universels". On suppose que les gènes "non-universels" ont été acquis par certaines souches par HGT, et qu'ils ne sont présents que de façon transitoire dans les génomes. En effet, un tel HGT est sûrement récent, et la probabilité qu'il soit fixé dans la population est très faible (plus précisément $1/N$ d'après Kimura (1962) s'il n'apporte pas d'avantage sélectif). Ainsi, la majorité de ces gènes ne seront pas fixés dans la population.

Le jeu de données que nous avons assemblé (espèces de départ et toutes leurs souches) est représenté par la figure 3.20. Malgré le nombre élevé de génomes disponibles, assembler un jeu de données équilibré n'est pas facile pour deux raisons : premièrement, plusieurs espèces avec le même nom de genre et d'espèce sont moins semblables entre elles que la plupart des souches des autres espèces. Soit ces "souches" sont en réalité des espèces différentes identifiées à tort comme faisant partie d'une même espèce, soit elles sont réellement dissemblables. Lors de la reconstruction d'un arbre, ces souches ont une longueur de branche les séparant plus grande. Dans notre cas, *H. pylori* est légèrement affecté par ce phénomène.

cinq espèces forment un sous-groupe monophylétique, une seule perte dans la branche menant à leur ancêtre commun suffit à expliquer la répartition du gène. Cependant, nous ne nous intéressons pas aux phylogénies intra-groupes.

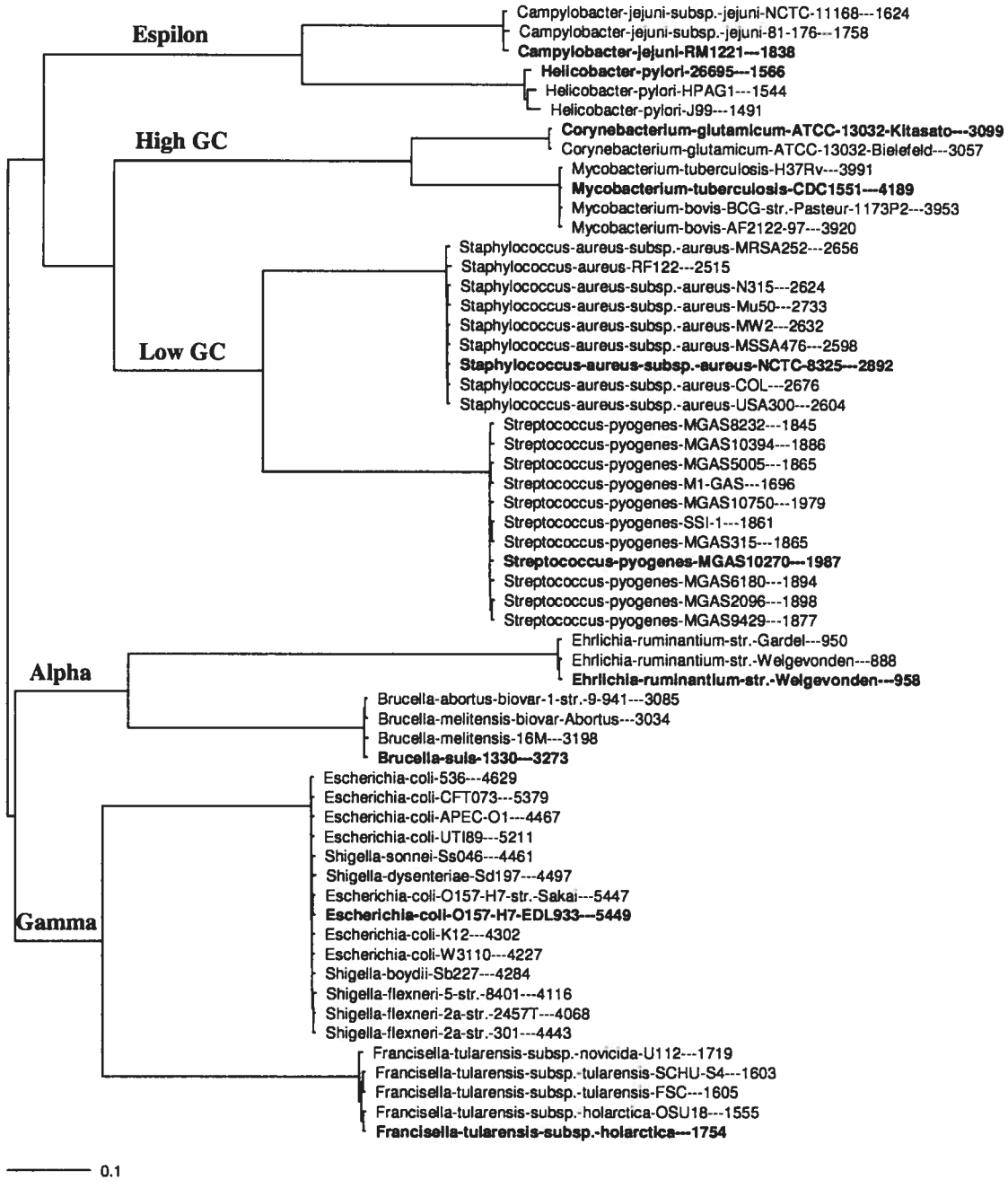


Figure 3.20 – 5x2 souches : arbre. Les espèces de départ sont en gras.

Deuxièmement, comme le montre notre arbre, les pathogènes humains font l'objet de plus d'efforts de séquençage et ont beaucoup de souches disponibles (exemple : *E. coli*, *S. aureus* et *S. pyogenes*), alors que d'autres espèces n'en ont que très peu. Nous verrons plus loin qu'un trop faible nombre de souches pour une espèce peut poser quelques problèmes.

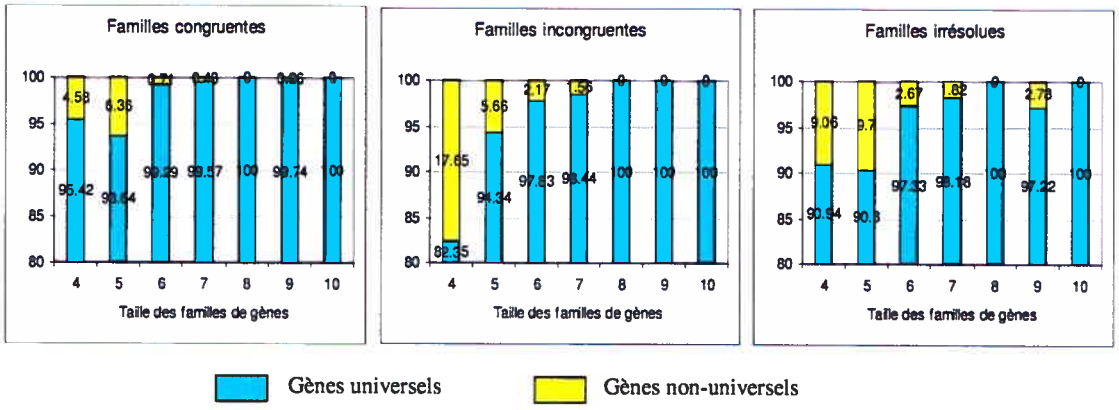


Figure 3.21 – 5x2 souches : Congruence, incongruence et irrésolution pour les gènes universels et non-universels. Noter le minimum inhabituel de 80% pour l'axe vertical.

L'analyse que nous réalisons ici consiste à déterminer le statut de chaque gène parmi les souches (universel ou non-universel), puis à corréler ces informations avec les données sur la congruence, l'incongruence et l'irrésolution des groupes que nous avons calculées précédemment. Les résultats sont présentés dans la figure 3.21. Les gènes non-universels font le plus souvent partie de familles de petite taille (4 et 5). De plus, la catégorie des familles incongruentes de taille 4 est celle qui contient le plus de gènes non-universels (17,65%). Ces résultats indiquent que le taux de HGT augmente quand les gènes sont absents des souches voisines. D'un point de vue biologique, cela signifie que ces gènes ont probablement été acquis récemment et indépendamment par certaines souches, mais pas par les autres. Or Hao et Golding (2004) ont constaté qu'au niveau des branches externes (de l'arbre des espèces), les insertions de gènes étrangers (par HGT donc) sont beaucoup plus nombreuses que les délétions. Les tailles des génomes bactériens étant globalement constantes (Kurland, 2005), cela implique que ces gènes récemment acquis seront

perdus sous peu[5]. Cependant, nous ne pouvons affirmer ceci avec certitude à cause de l'échantillonnage limité de certaines souches dans notre jeu de données. En effet, si une espèce donnée n'a que deux souches, seuls deux états sont possibles : considérant que le gène est présent chez la souche de départ, il peut être soit présent soit absent chez l'autre souche. Le pourcentage de présence du gène est donc soit 50% (non-universel) ou 100% (universel). La présence de plusieurs souches par espèce permet d'avoir des pourcentages de présence plus fins, et autorise la création de seuils pour mieux quantifier la non-universalité. Par exemple, avec 5 souches ou plus, on pourrait définir une catégorie de 80% de présence, ce qui donnerait plus de sensibilité à notre analyse.

Un autre fait mérite d'être mentionné. Lors de la détection par blast dans les souches voisines, nous obtenons parfois des matchs significatifs[6] multiples au sein d'un même génome. Les matchs multiples révèlent la présence de paralogues proches pour notre gène de départ. Nous n'avons pas utilisé cette information car elle ne nous était pas nécessaire, mais on pourrait imaginer une analyse supplémentaire qui consisterait à corréler la présence de paralogues avec les taux de HGT. Nous postulons que cette corrélation serait positive : en effet, compte tenu de la faible occurence des duplications de gènes ches les procaryotes, les paralogues aurait été acquis par HGT.

## 3.13 Application à une échelle évolutive restreinte

Les choix d'espèces que nous avons réalisés jusqu'à présent étaient à l'échelle des Bactéries, c'est-à-dire avec des espèces des grands groupes bactériens. Le but de la présente manipulation est de vérifier que notre hypothèse, selon laquelle les gènes peu répandus sont plus transférés que les gènes universels, est aussi vérifiée à une échelle évolutive plus réduite. Nous appliquons donc les mêmes critères que précédemment pour la sélection des espèces (voir section 2.1.2) mais en nous restreignant à un seul groupe. Notre choix s'est porté sur les gamma-protéobactéries,

[5] à l'échelle évolutive

[6] Notre seuil de significativité est de 90% d'identité au niveau des acides aminés.

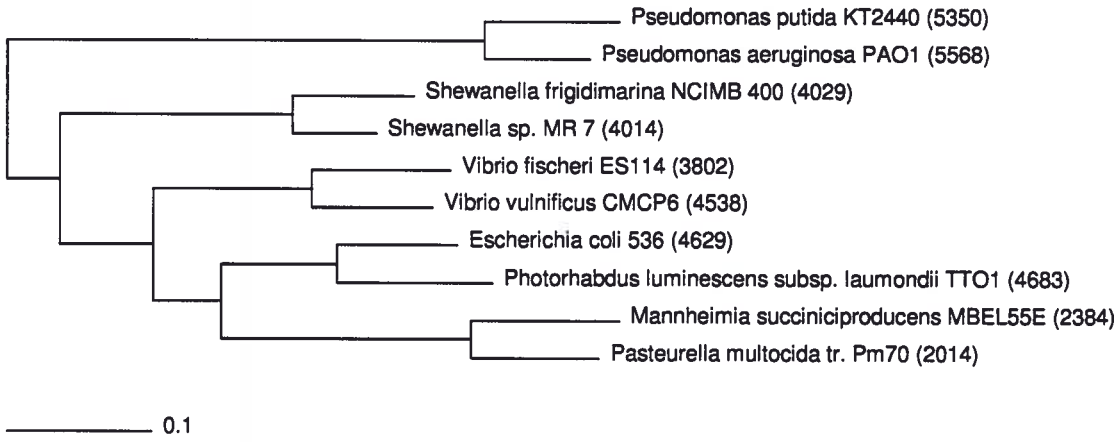


Figure 3.22 – 5x2 gamma : arbre. Toutes les espèces sont des gamma-protéobactéries.

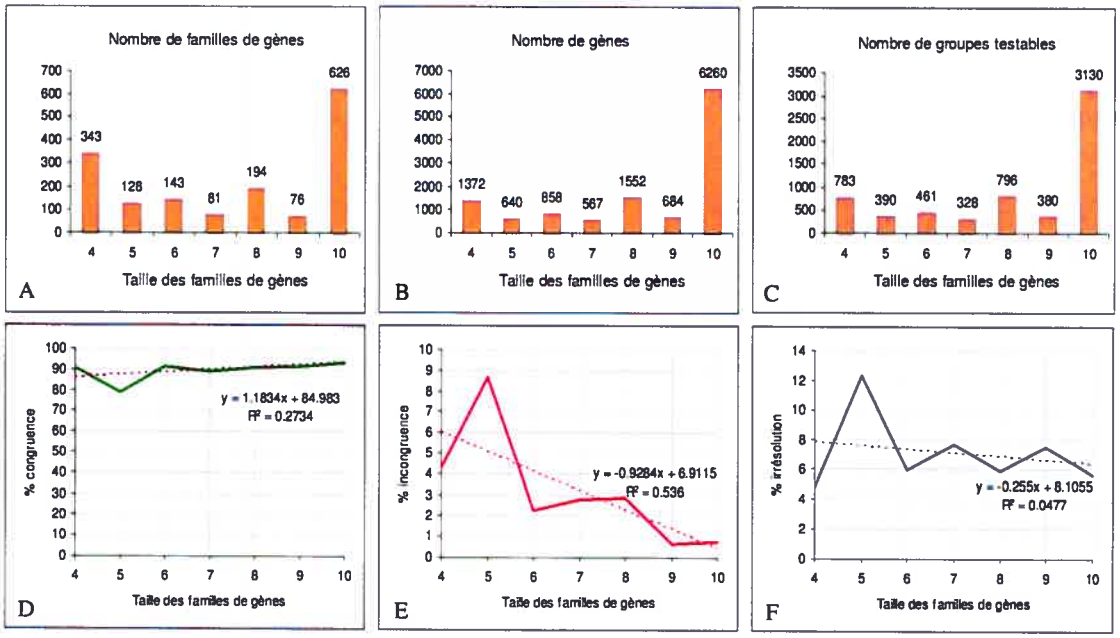


Figure 3.23 – 5x2 gamma 1e-4 : congruence, incongruence, irrésolution, effectifs.

car elles comptent le plus grand nombre d'espèces séquencées. L'arbre des espèces est représenté dans la figure 3.22. Les effectifs obtenus pour chaque taille de famille ont le même profil que dans nos jeux de données classiques, à savoir que les tailles avec le plus de familles sont 4 et "tous" (10 dans ce cas) (voir figure 3.23 A). Par contre, il y a comparativement moins de familles de taille 4 (343), et plus de taille 10 (626) que pour notre jeu de données 5x2 original (voir figure 3.7) dont les effectifs étaient 225 et 133 respectivement. Ceci est certainement la conséquence de la faible distance évolutive entre les gamma-protéobactéries : elles ont un grand nombre de gènes en commun, et comparativement moins de gènes sont spécifiques à des sous-groupes de gamma-protéobactéries qu'à des groupes bactériens entiers (5x2 original). Les effectifs des tailles intermédiaires sont également plus faibles que les tailles 4 et 10, mais les familles de tailles paires sont plus nombreuses que celles de tailles impaires, comme c'était le cas pour le jeu 5x2 original. Enfin, on remarquera que le nombre global de familles est plus élevé, ce qui est évidemment dû à la proximité phylogénétique.

Les valeurs de congruence, incongruence et irrésolution (figure 3.23, D, E et F) présentent également le même profil que le jeu de données original. Le taux de congruence est cependant plus élevé, autour de 90% (par rapport à environ 80% auparavant). Le taux d'incongruence présente le même profil avec une pente négative, le pourcentage brut étant cependant plus faible. La différence quantitative de ces taux s'explique, comme pour les effectifs, par la proximité phylogénétique des espèces. Leur similarité qualitative (c'est-à-dire au niveau de la pente) renforce notre hypothèse que les gènes rares sont plus transférés que les gènes répandus.

Le nombre important de familles de gènes détectées avec le seuil standard de 1e-4 nous a autorisé à conduire une expérience avec un seuil beaucoup plus sévère de 1e-100. Les résultats (voir figure 3.24) montrent que le taux de congruence (D) est extrêmement élevé, étant supérieur à 95% (sauf pour la taille 5). De même, le taux d'incongruence (E) est inférieur à 2%, sauf pour la taille 5. Enfin, l'irrésolution (F) est elle aussi très réduite, à moins de 3% (sauf pour les tailles 5 et 7 dont nous discuterons plus loin), ce qui est nouveau, car elle était de l'ordre de 10% avec les

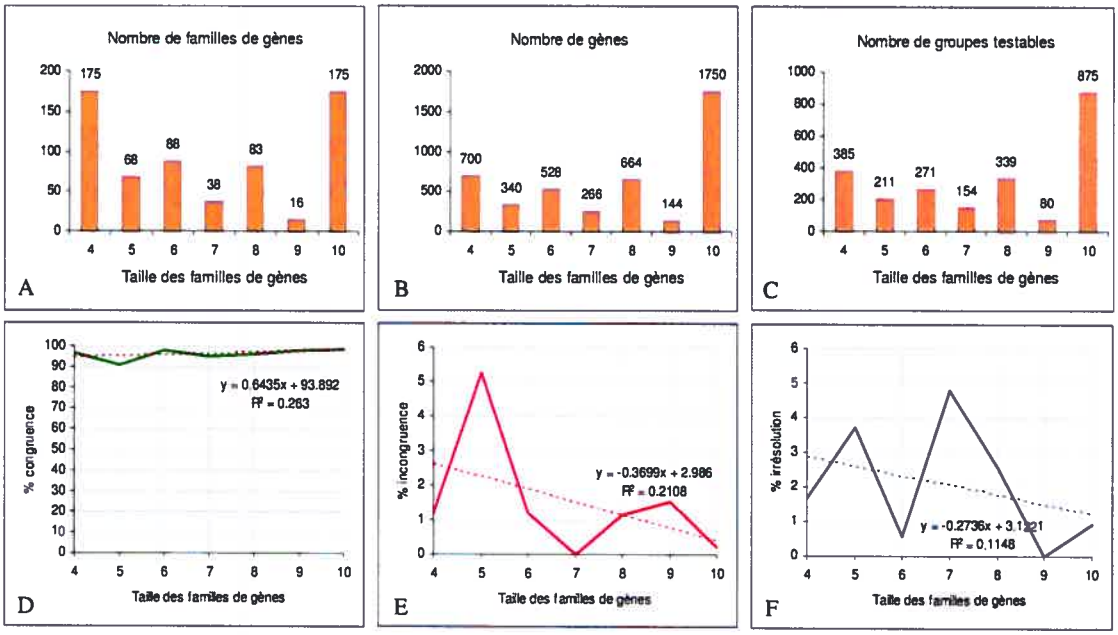


Figure 3.24 – 5x2 gamma 1e-100 : congruence, incongruence, irrésolution, effectifs

jeux de données classiques. Ces résultats sont causés par la valeur extrême du seuil de brh. Comme nous l'avons vu à la section 3.10, les familles dont la e-value était inférieure à 1e-40 avaient une congruence légèrement supérieure et une incongruence légèrement inférieure à celles des familles avec une e-value supérieure à 1e-40. Dans notre cas, toutes nos familles ont une e-value inférieure à 1e-100, et les résultats vont dans le même sens, mais sont plus marqués. De plus, la longueur moyenne

| Taille | 1e-100 | 1e-4 |
|---|---|---|
| 4 | 430 | 226 |
| 5 | 465 | 267 |
| 6 | 456 | 262 |
| 7 | 490 | 284 |
| 8 | 463 | 279 |
| 9 | 555 | 291 |
| 10 | 558 | 315 |
| Moyenne | 488 | 275 |

Tableau 3.4 – 5x2 gamma : longueurs moyennes des familles (acides aminés après sélection par GBLOCKS) pour 1e-100 et 1e-4

des gènes (voir tableau 3.4) donne aussi le même résultat que précédemment : pour 1e-100, les gènes ont en moyenne 448 acides aminés contre seulement 275 pour 1e-4[7]. Ainsi, il semble y avoir une corrélation positive entre l'e-value d'une famille, la longueur de l'alignement et le taux de congruence. La corrélation est négative avec l'incongruence et l'irrésolution. Cependant, nous n'avons pas d'explication satisfaisante à ce phénomène, et il faudrait concevoir des expériences spécifiques afin de l'étudier.

Discutons à présent du cas des familles de taille 5 et 7. Nous savons par l'expérience sur la conformation des familles (voir section 3.11) que les familles de tailles impaires[8] contiennent plus de gènes transférés du fait de la présence accrue de singletons. Or la famille 7 présente un taux d'incongruence de 0, ce qui est tout à fait inattendu. Par contre, le taux d'irrésolution forme un pic à presque 5%. La taille 5 présente elle aussi un pic d'irrésolution, et ce en plus de son pic d'incongruence. Comme il est très improbable qu'il n'y ait pas de transfert parmi les familles de taille 7[9], cela renforce l'hypothèse de l'irrésolution comme indicateur de HGT que nous avons formulée plus tôt (voir figure 3.4). Cependant, l'irrésolution n'est que légèrement supérieure à l'incongruence, leur proportion ayant le même ordre de grandeur que pour le jeu 5x2 original. Étant donné que nous n'avons échantillonné qu'un seul groupe taxonomique, une plus grande portion des HGT devrait provenir de "l'extérieur", et donc se traduire par une irrésolution nettement supérieure à l'incongruence. Or ce n'est pas le cas ici. Deux explications non exclusives existent. La première est que la part de l'irrésolution dénotant les HGT est peut-être faible par rapport à celle résultant du signal non-phylogénétique. La seconde découle du fait que nous sommes capables de détecter les HGT à l'intérieur des gamma-protéobactéries puisqu'elles sont divisées en sous-groupes, ce qui n'était pas le cas auparavant. Comme il est vraisemblable que les gamma-protéobactéries échangent aisément des

[7]Dans ce cas-ci, et contrairement à la section 3.10, les familles à 1e-100 sont incluses dans celles à 1e-4.

[8]dans les jeux de données avec 2 espèces par groupe.

[9]Surtout à cause de la présence de familles avec 1 et 3 singletons, comme nous l'avons montré à la section 3.11.

gènes[10], l'incongruence observée pour ce jeu de données serait principalement le résultat des transferts à l'intérieur de ce groupe.

Si l'on réexamine nos résultats d'incongruence précédents, on remarque des pics similaires, quoique moins marqués. C'est le cas notamment pour les figures 3.15 (rééchantillonnage 5x2 : pics à taille 7 et 9) et 3.16 (figure du bas, familles avec e-values supérieure à 1e-40 : pics à 5, 7 et 9).

Enfin, nous avons réalisé l'étude des singletons pour les deux seuils de 1e-4 et 1e-100. Les résultats (figure 3.25) confirment ceux obtenus précédemment (voir section 3.11). Le taux d'incongruence augmente avec le nombre de singletons. La seule anomalie à 1e-4 est une congruence à 3 singletons supérieure à celle pour 2, et une incongruence nulle. Le nombre réduit de familles (13) peut expliquer ce résultat aberrant. Ainsi, les principaux résultats que nous avons obtenus avec des

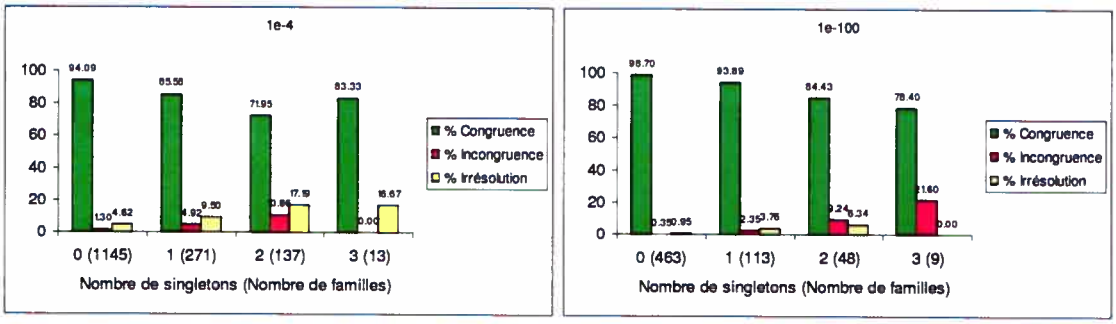


Figure 3.25 – 5x2 gamma 1e-4 et 1e-100 : singletons.

jeux de données à l'échelle des bactéries sont globalement retrouvés au sein des gamma-protéobactéries. Il serait intéressant d'étendre ce test à d'autres groupes, mais le nombre et/ou la diversité des espèces disponible n'est pas encore suffisante à cet effet.

## 3.14 Autres configurations de l'échantillonnage taxonomique

Nous avons jusqu'ici volontairement limité les configurations de groupes et d'espèces dans nos jeux de données à 5x7 et 5x2 afin de ne pas induire de variation

[10]En effet, si les HGT peuvent se produire entre des espèces éloignées, la proximité phylogénétique engendre des environnements génomiques plus similaires favorisant la fixation des HGT.

pour ces deux paramètres. À présent nous introduisons deux nouveaux types : 12 groupes × 2 espèces et 6 groupes × 4 espèces.

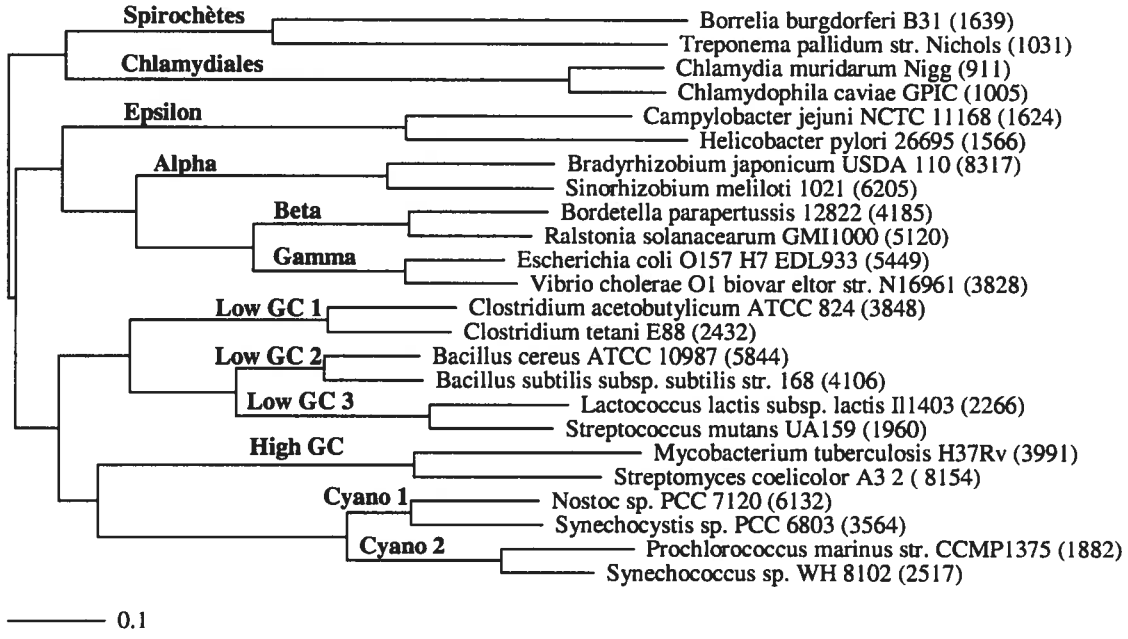


Figure 3.26 – 12x2 : arbre

L'arbre 12x2 (figure 3.26) présente des branches inter-groupes plus courtes que souhaité par notre protocole (Beta, Gamma, Low GC 2, Cyanobactéries 1 et 2 surtout). Nous avons dû faire un tel compromis afin de pouvoir assembler autant de groupes. Malgré cela, les résultats corroborent ceux obtenus précédemment. Les effectifs (figure 3.27 A, B et C) présentent la même distribution bimodale centrée sur les familles de petites tailles (4 à 6) et universelles (24). Les tendances des taux de congruence (D) et d'incongruence (E) sont également similaires à précédemment : la congruence augmente avec la taille des familles alors que l'incongruence diminue. Par contre, l'irrésolution (F) augmente légèrement, traduisant la difficulté à bien résoudre tous les groupes, et ce à cause du manque de signal phylogénétique dans certaines branches de l'arbre (les branches trop courtes). Il y a également de nombreux pics d'incongruence et d'irrésolution, mais on ne retrouve pas leur association avec les tailles impaires, bien que nous soyons dans une configuration avec 2 espèces par groupe. On croit discerner une périodicité de 3 (tailles 7, 10, 13 pour

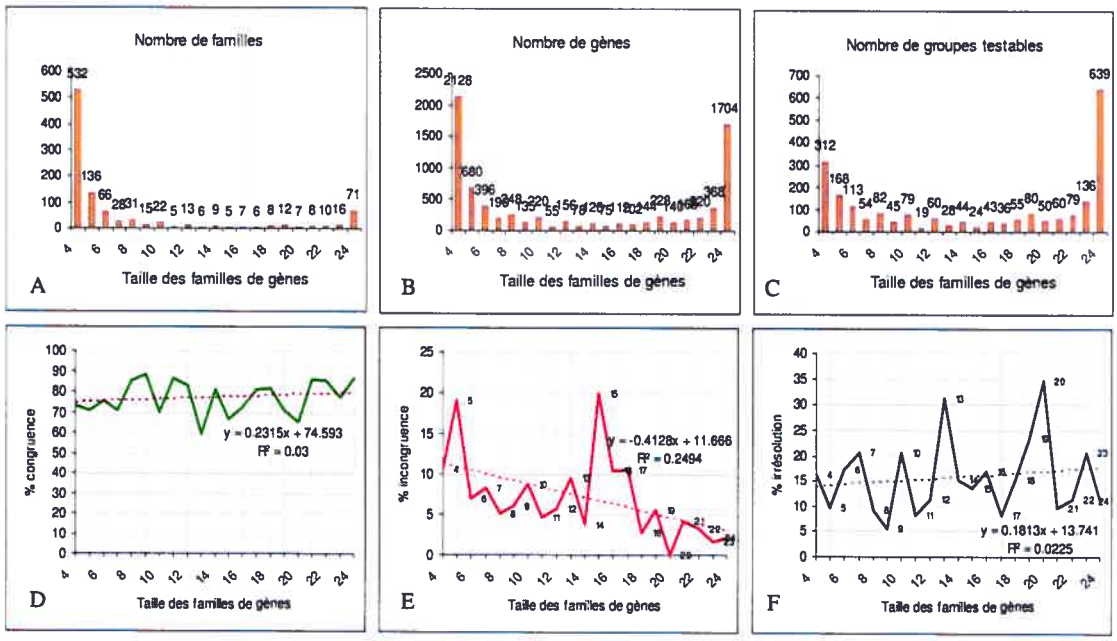


Figure 3.27 – 12x2 : A, B, C : effectifs des familles, des gènes et des groupes testables. D, E, F : congruence, incongruence et irrésolution. Les tailles des familles sont indiquées sur la courbe.

E et F), mais les effectifs de groupes étant relativement faibles, on n'est pas à l'abri d'erreurs stochastiques, comme dans le cas 5x7 (voir section 3.1 ; cependant les effectifs plus nombreux cette fois-ci). Dans ce jeu de données, on vérifie bien que les gènes rares sont plus sujets aux HGT que les gènes plus répandus, vérifiant d'une part notre hypothèse, et montrant d'autre part que notre protocole est robuste à des violations des conditions de départ puisque certaines branches inter-groupes courtes ne l'ont pas empêché de détecter une baisse nette de l'incongruence (et donc des HGT) en fonction de la diffusion des gènes.

L'arbre 6x4 (figure 3.28) montre que ce jeu de données est plus conforme à nos principes de sélection des espèces que le 12x2, toutes les branches inter-groupes étant suffisamment longues cette fois. Sans surprise, les effectifs des familles de tailles intermédiaires (figure 3.29 A, B et C) sont faibles. À nouveau, la congruence (D) augmente avec la diffusion des gènes, alors qu'incongruence et irrésolution diminuent. Comme pour 12x2, les pics d'incongruence et d'irrésolution ne sont pas

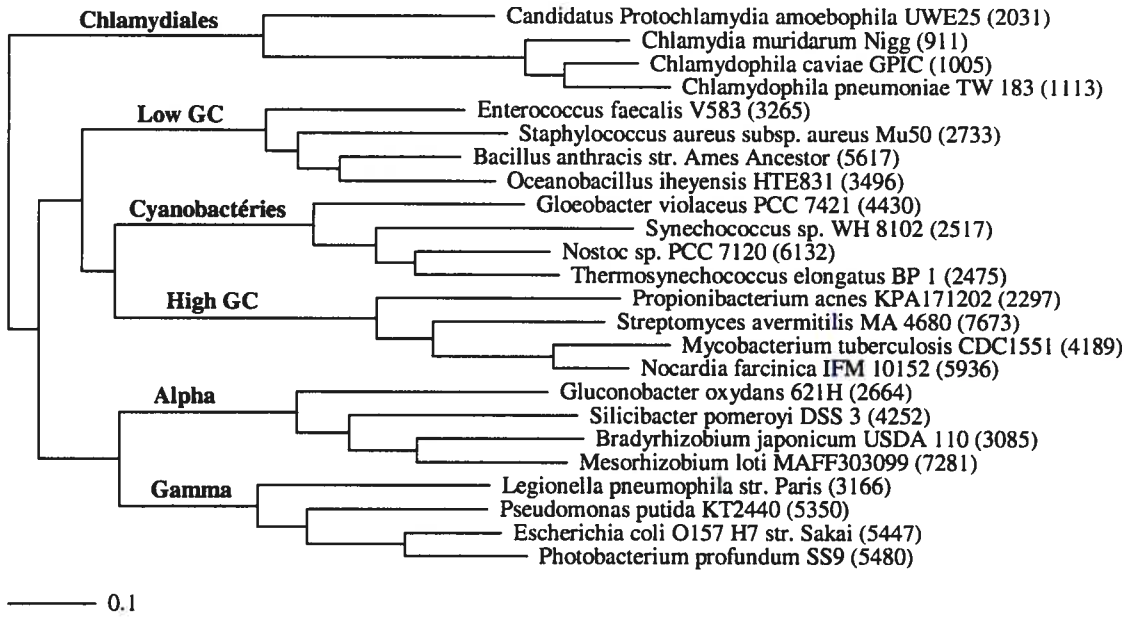


Figure 3.28 – 6x4 : arbre

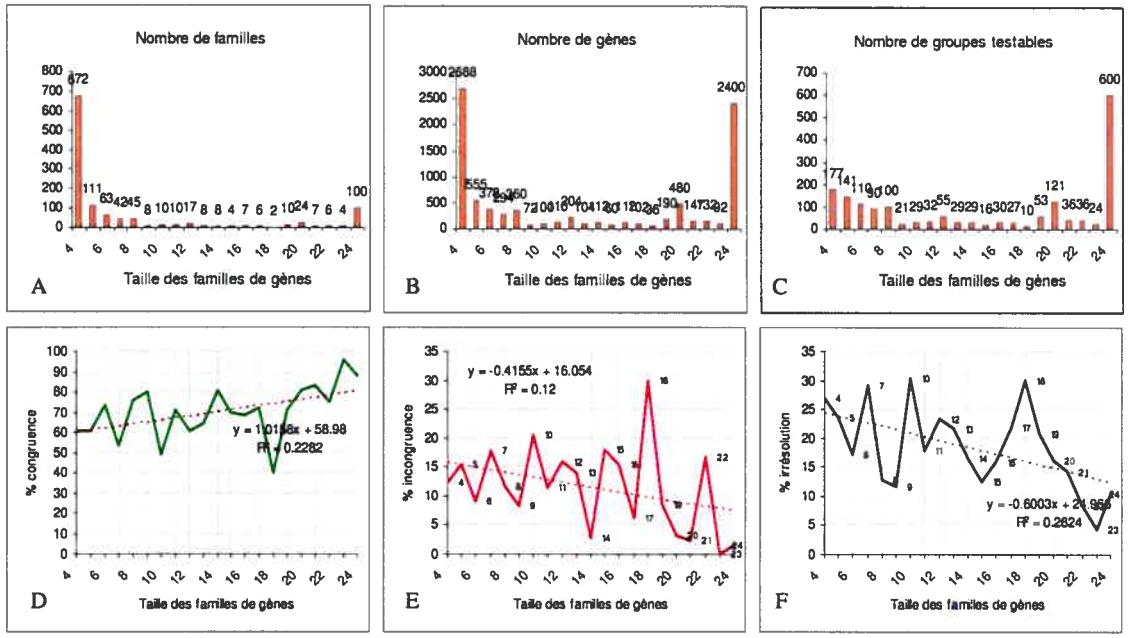


Figure 3.29 – 6x4 : A, B, C : effectifs des familles, des gènes et des groupes testables. D, E, F : congruence, incongruence et irrésolution. Les tailles des familles sont indiquées sur la courbe.

exclusivement associés à des tailles de familles impaires[11].

Bien que les effectifs des tailles intermédiaires des jeux de données 12x2 et 6x4 soient plus faibles que pour les jeux 5x2 à cause du nombre plus élevé d'espèces (24 dans les deux cas), les résultats de congruence, d'incongruence et d'irrésolution confirment ceux obtenus avec les jeux de données de type 5x2 et 5x7. Notre protocole apparaît donc en mesure de montrer la validité de notre hypothèse initiale quel que soit le type de conformation des jeux de données en termes de nombre de groupes et d'espèces par groupe.

## 3.15 Création de jeux plus petits à partir de 5x7

Les différentes configurations de jeux de données ne peuvent être comparées que jusqu'à un certain point, car les espèces sont variables d'un jeu à l'autre. Dans cette section nous partons d'un jeu de données riche en espèces (notre jeu 5x7 initial, voir 3.1), puis nous retirons successivement une espèce à l'ensemble des groupes pour aboutir à un jeu 5x2. De cette façon nous pouvons réellement comparer l'effet du nombre d'espèces par groupe sur la sensibilité de notre méthode. Pour chaque jeu de données intermédiaire, les arbres phylogénétiques sont recalculés. D'autre part, nous re-détectons aussi les familles de HNP parmi le nouveau jeu d'espèces, avant de reconstruire les arbres. Les résultats obtenus sont présentés à la figure 3.30. Les résultats avec la re-détection des HNP sont semblables à ceux obtenus sans. Non seulement les courbes de régression sont presque égales, mais les courbes sont très souvent superposées, y compris lorsqu'il y a des pics importants. Cela signifie que les HNP détectés par `brh` sont peu influencés par le nombre d'espèces présentes. Autrement dit, notre méthode de détection des HNP est robuste. La seconde observation est que les jeux de données les plus sous-échantillonnés (les plus réduits) sont ceux qui contiennent le moins de HGT : les niveaux d'incongruence et d'irrésolution diminuent effectivement avec le nombre d'espèces par groupe. On pourrait arguer que ceci n'est qu'un artéfact causé par le fait que des arbres de petite taille

[11] Ce à quoi on pourrait s'attendre comme pour les cas avec 2 espèces par groupe, puisqu'on a $2 \times 2$ ici.

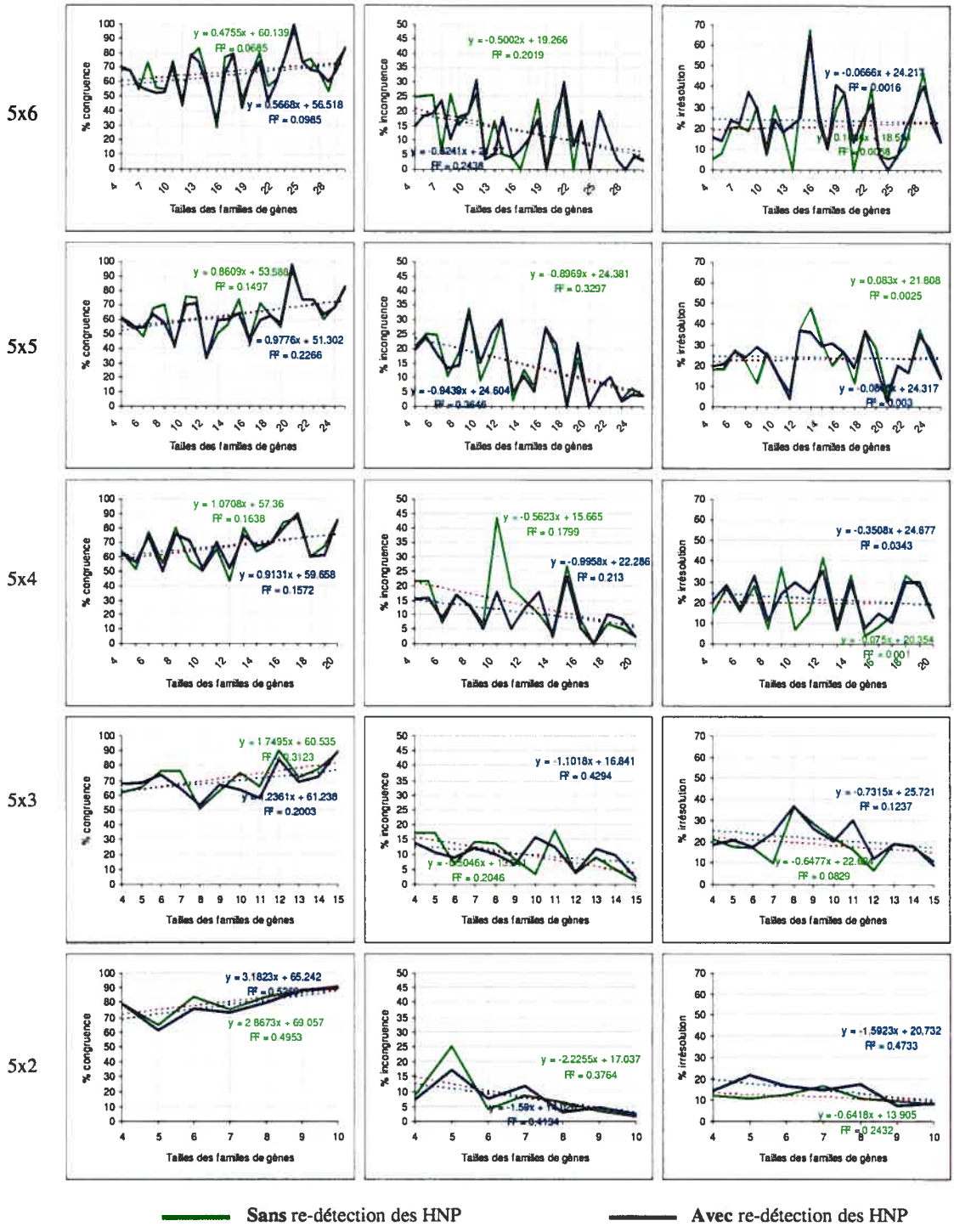


Figure 3.30 – 5x7 : downsampling. À partir du jeu de données 5x7, on retire une espèce de chaque groupe pour obtenir le jeu 5x6. Le processus est répété pour les autres jeux. Les échelles des axes verticaux ont été maintenus constants afin de favoriser les comparaisons.

sont plus faciles à inférer que les grands ; la reconstruction de grands arbres donnerait donc plus de contradictions et d'indécisions (incongruences et irrésolution). Cependant, une autre explication serait la suivante : plus il y a d'espèces, plus la probabilité d'avoir des HGT est élevée. Or notre protocole est conçu de telle manière qu'une seule espèce ayant reçu un gène d'un autre groupe rendra son propre groupe incongruent, et ce peu importe le nombre d'espèces. La sensibilité de notre méthode augmente donc avec le nombre d'espèces par groupe.

## 3.16 Tirage aléatoire d'un grand nombre d'espèces

Tous nos résultats ont été obtenus avec un nombre relativement petit de jeux de données. De plus, ces derniers ne sont pas totalement indépendants car certaines espèces sont utilisées dans plusieurs configurations, souvent à cause de leur position taxonomique favorable et lorsqu'il y a peu de représentants de leur groupe. Afin de montrer que nos résultats ne sont pas le fruit de particularités dues aux choix des espèces, nous créons un bassin d'espèces pour 5 groupes[12]. Puis nous choisissons aléatoirement 2 espèces dans chacun afin de constituer un jeu 5x2. Le processus est répété 100 fois.

Les résultats (voir figure 3.31) confirment les grandes tendances déjà observées. Tout d'abord, on retrouve la légère pente descendante de l'incongruence en fonction de la diffusion des gènes, ce qui appuie notre hypothèse. Les pics d'incongruence pour les valeurs impaires sont clairement présents, confirmant l'association entre singletons et HGT. L'artéfact de la taille 4 qui cause une surestimation de la congruence des familles de cette taille est aussi nettement visible : la congruence à 4 est augmentée alors que l'incongruence (et dans une moindre mesure l'irrésolution) est sous-estimée. On constate que certains groupes s'écartent nettement de l'incongruence moyenne. D'une part les cyanobactéries ont le taux d'incongruence le plus faible parmi les 5 groupes testés. De plus elles ont aussi le plus faible taux d'incongruence. On peut donc affirmer qu'elle ont un faible taux de HGT, ce qui

[12] Cyanobactéries : 11 espèces, High GC : 16, Low GC : 24, Gamma-protéobactéries : 35, Alpha-protéobactéries : 20.

confirme la tendance observée pour notre jeu de données 5x2 original (voir section 3.3). D'autre part, les gamma-protéobactéries ont un taux d'incongruence élevé, indiquant une propension aux HGT. Ceci est en accord avec la grande diversité écologique de ce groupe, car les HGT participent grandement à l'adaptation des espèces à de nouvelles niches. Par contre, comme le nombre de spirochètes disponibles est trop faible, nous ne pouvons confirmer le haut taux de HGT observé pour ce groupe dans notre jeu 5x2 original. Enfin, comme les résultats obtenus par une méthode rapide mais sujette aux artéfacts comme la parcimonie (MP) sont comparables à ceux d'une méthode robuste comme le maximum de vraisemblance (ML), notre choix d'utiliser une méthode de distance (relativement rapide) s'est avéré valable tel que nous l'avions postulé.

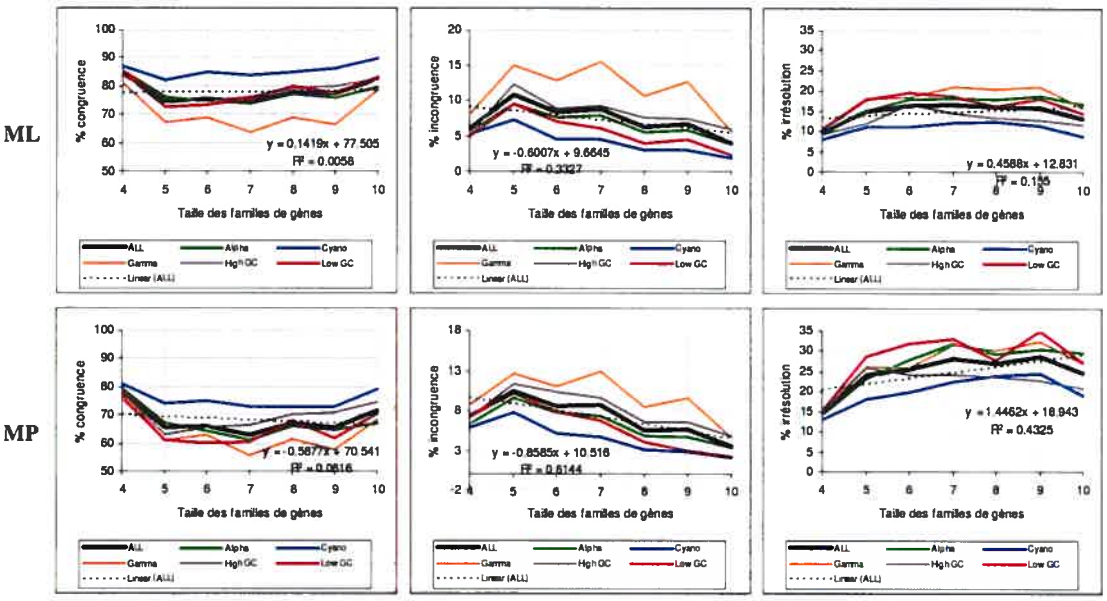


Figure 3.31 – Tirage de 100 jeux de données 5x2 : congruence, incongruence et irrésolution en ML (en haut) et en MP (en bas).

# CHAPITRE 4

# CONCLUSION ET PERSPECTIVES

## 4.1 Notre hypothèse est vérifiée

L'hypothèse que nous avons formulée initialement était la suivante : les gènes répandus sont moins sujets au transfert horizontal (HGT) que les gènes avec une distribution taxonomique plus limitée. Un gène arrivant par transfert aurait très peu de chance d'être fixé dans le génome hôte si celui-ci possède déjà un gène orthologue remplissant la même fonction. Un gène universel étant déjà présent chez toutes les espèces, un transfert réussi serait improbable. En revanche, un gène rare a plus de chances d'apporter une fonction nouvelle à l'espèce hôte, lui apportant possiblement un avantage sélectif qui conduira à la fixation du nouveau gène dans la population.

Les résultats ont confirmé que notre hypothèse est vraie : dans tous nos jeux de données, l'incongruence présentait une pente négative vis-à-vis de la répartition taxonomique. Comme notre protocole (notamment le choix d'espèces) est conçu afin d'éliminer les artéfacts de reconstruction grâce à un signal phylogénétique (SP) très supérieur au signal non-phylogénétique (SNP), nous avons attribué l'incongruence aux HGT. L'incongruence des gènes rares étant plus élevée, nous en déduisons qu'ils sont plus sujets aux HGT que les gènes "universels" (présents chez toutes les espèces du jeu de données). Cependant, la pente est faible.

Par ailleurs, plusieurs indices nous ont amenés à considérer l'irrésolution comme un autre indicateur de HGT. Tout d'abord, malgré le contexte d'une inférence phylogénétique facile, l'irrésolution était relativement élevée. D'autre part, elle suivait souvent la même tendance à la baisse que l'incongruence pour les familles de grande taille, alors qu'en principe on s'attendrait à ce qu'elle augmente en raison du plus grand nombre de noeuds à résoudre. Nous avons donc associé une partie de l'irrésolution aux HGT ayant pour source un groupe donneur non représenté dans nos

jeux de données. Ceci s'explique par le fait que le gène transféré ne ressemblant à aucun autre dans le jeu de données, il est attiré vers la base de l'arbre, entraînant l'irrésolution.

Ainsi, la diminution des HGT avec la hausse de la répartition est donc vraisemblablement sous-estimée, mais elle n'est probablement pas la première cause dans la fixation des HGT.

## 4.2 Évaluation de notre protocole

Notre protocole s'est révélé robuste à la variation des configurations des jeux de données, donnant des résultats cohérents pour 5 à 12 groupes, avec 2 à 7 espèces par groupe. Cependant, les jeux de données avec un nombre élevé d'espèces ont des effectifs pour les familles de taille intermédiaire très faibles, ce qui induit des erreurs stochastiques. Afin d'y remédier, nous avons assoupli le critère de complétude des sous-graphes de brh, ce qui a légèrement contribué à augmenter le nombre de familles détectées. Nous avons devisé d'autres solutions possibles. Par exemple, la combinaison de deux jeux de données ou plus *après* la détection des familles d'homologues non-paralogues (HNP) : on doublerait ainsi facilement les effectifs des familles. Par exemple, au lieu de prendre 6 espèces par groupes, on pourrait créer deux sous-jeux de données avec 2 espèces par groupe en utilisant les mêmes espèces. On s'affranchirait ainsi de la sévérité du critère de complétude des brh car chaque graphe nécessiterait comparativement beaucoup moins de brh (voir figure 4.1). Comme les jeux de données sont disjoints, ils sont indépendants, et les familles détectées peuvent être traitées ensemble. Par contre, il faudrait trouver un moyen d'assembler les sous-familles afin de reconstituer les familles d'origine. Ceci pourrait être fait en cherchant quelques brh entre les sous-familles.

Une autre façon de détecter plus de familles serait d'examiner plus en détail les familles rejetées par notre programme de détection `brh` afin d'en "repêcher" quelques unes qui satisferaient à certains critères. Par exemple, il se pourrait qu'un seul des gènes ait un brh vers un gène extérieur à la famille pour que celle-ci soit

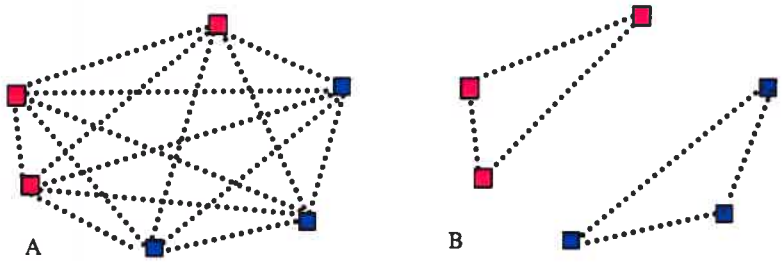


Figure 4.1 – A. Jeu de données original : 15 brh sont nécessaires pour sélectionner une famille de taille 6. B. La division en deux jeux de données diminue drastiquement le nombre de brh nécessaires pour détecter les mêmes 6 gènes.

rejetée. On pourrait choisir de la conserver en excluant le gène connecté par un seul brh.

Une élégante hypothèse pour expliquer la bimodalité de la distribution des familles (et donc le petit nombre de familles de taille intermédiare) a été proposée par Ignacio G. Bravo. Les gènes rares seraient en train de se répandre parmi les espèces, alors que les gènes universels seraient en voie de perte chez certaines espèces (voir figure 4.2).

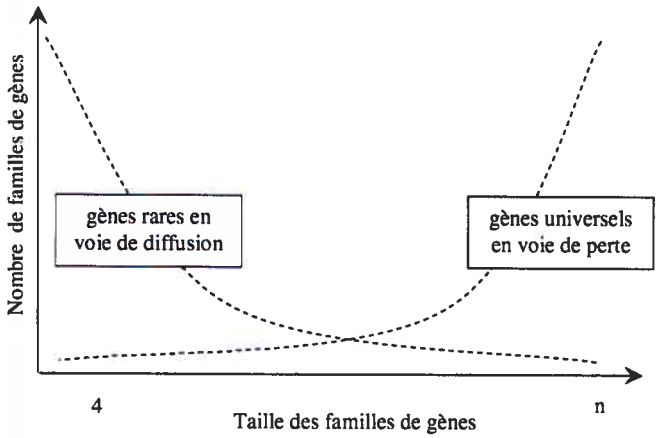


Figure 4.2 – Hypothèse de Bravo

## 4.3 Proportion des gènes détectés par `brh` et leur représentativité

Bien que nous ayons abondamment analysé les effectifs des familles de HNP détectées par **brh**, nous n'avons jamais analysé leur nombre par rapport au nombre total de gènes des espèces en question. C'est ce que nous nous proposons de faire à présent. Le tableau 4.1 montre que la proportion de gènes détectés par **brh** est étonnamment constante à travers tous nos jeux de données, variant de 8,71% pour le jeu 12x2 à 9,99% pour 5x2. On conviendra que c'est une part modeste

| | | | | | |
|---|---|---|---|---|---|
| 5x7 | 9,55% | 6x4 | 9,43% | 5x2 souches | 9,24% |
| 5x2 | 9,99% | 12x2 | 8,71% | 5x2 gamma | 29,10% |

Tableau 4.1 – Proportion des gènes détectés par **brh** pour nos jeux de données avec un seuil de 1e-4. (nombre de gènes détectés par **brh** / nombre total de gènes dans les espèces)

mais non négligeable des génomes complets. La seule exception est pour le jeu 5x2 gamma, pour lequel les gènes détectés par **brh** représentent 29,1% de la taille totale des génomes. Ce résultat est évidemment causé par la proximité phylogénétique des espèces du jeu de données (des gamma-protéobactéries). Comme les résultats obtenus pour ce jeu n'étaient pas différents des autres, cela indique que les 10% de gènes détectés dans les jeux de données "normaux" ont le même profil de HGT que 30% des gènes. Par contre, toute extrapolation aux 70% de gènes restants doit être faite avec la plus grande prudence.

## 4.4 Gènes non détectés par `brh`

Puisque nous ne détectons que 10 à 30% des gènes, quels sont les gènes que nous ne détectons pas ? Ces gènes se classent en trois catégories. (1) Tous les gènes ayant au moins un paralogue : comme l'explique la figure 2.3, toute famille d'homologues non-paralogues dont au moins un des membres a un paralogue proche sera rejetée. Comme le nombre de paralogues n'est pas négligeable, ils entrainent la mise à l'écart d'un nombre probablement important de gènes. (2) Les gènes très rares : le nombre minimum d'espèces pour avoir une phylogénie est de 4. Nous ignorons donc

les gènes présents chez trois espèces ou moins. Techniquement, `brh` est capable de détecter les gènes présents chez 3 et 2 espèces, mais pas 1 car il faut qu'il y ait au moins un brh. (3) Les orthologues issus de fusion ou de fission : ces gènes auront seulement une portion de similarité avec les autres orthologues, ce qui empêche la réciprocité du hit BLAST, et ainsi leur détection par `brh`. Enfin, en plus de ces trois catégories de gènes, il existe dans chaque génome un certain nombre d'ORF n'ayant pas d'homologue connu.

## 4.5 Amélioration de la détection des HGT

Plusieurs raffinements peuvent être mis en place afin d'améliorer la détection des HGT. Un jeu de données avec un nombre élevé d'espèces par groupe a une plus grande sensibilité aux HGT, mais au prix d'une saturation plus rapide. En effet, plus il y a d'espèces, plus il y a de chances qu'un transfert dans le jeu de données ait impliqué une des espèces. Un gène transféré (appartenant à une espèce d'un autre groupe) a donc plus de chances d'être groupé avec son gène d'origine, ce qui brise la monophylie du groupe.

Afin de détecter à la fois les transferts inter-groupes et intra-groupes, on pourrait imaginer la création de sous-groupes d'espèces parmi les groupes principaux. Cela reviendrait à créer plusieurs jeux de données comme 5x2 gamma (voir section 3.13) avec d'autres groupes taxonomiques (par exemple les Cyanobactéries), puis les à assembler dans un super-jeu de données. Cependant, on risque d'être limité par le nombre moindre d'espèces disponibles en dehors des gamma-protéobactéries.

Nous avons expliqué plus tôt que les transferts provenant de groupes non présents dans le jeu de données engendraient de l'irrésolution. Une solution pour convertir cette irrésolution en incongruence serait d'ajouter au jeu de données plusieurs groupes additionnels, mais représentés par une seule espèce à chaque fois afin de ne pas trop augmenter le nombre total d'espèces, ce qui, comme nous l'avons vu, peut diminuer les effectifs des familles de tailles intermédiaires. Ces espèces de groupes supplémentaires permettraient aux séquences transférées de se grouper

avec elles (et donnant ainsi de l'incongruence) plutôt que d'être attirées vers la base de l'arbre, qui se traduirait par de l'irrésolution.

## 4.6 Amélioration des analyses

Les analyses que nous avons menées sont loin d'être exhaustives. Mis à part l'expérience dans laquelle nous avons tiré 100 combinaisons d'espèces 5x2 (voir sections 2.2.13 et 3.16), nous n'avons pas réalisé de tirages aléatoires. L'expérience du retrait d'espèces (voir sections 2.2.12 et 3.15) en bénéficierait grandement : en effet, nous avons simplement retiré une à une les espèces choisies arbitrairement. Une répétition aléatoire du retrait aurait d'une part modifié l'ordre de retrait des espèces, et d'autre part changé les deux espèces restantes à la fin des retraits.

Maintenant que nous avons un protocole qui détecte de manière fiable les HGT parmi des génomes, nous sommes en mesure de vérifier l'hypothèse de la complexité (Jain et al., 1999). Selon cette hypothèse, les gènes *informationnels*, impliqués dans des interactions plus complexes, sont moins aisément transférés que les gènes *opérationnels* qui ont un nombre réduit des partenaires (voir section 1.5.1.3). Il suffit pour cela d'obtenir le nombre d'interactions des gènes, puis de croiser cette information avec le taux de HGT. Ainsi, on pourrait directement observer s'il y a une corrélation (positive) entre nombre d'interactions et taux de transfert. Une alternative consisterait à analyser les fonctions des gènes transférés, puis de déterminer s'il y a des catégories plus transférées que d'autres.

Avec notre méthode de détection des HGT, nous pouvons à présent étudier l'influence de divers paramètres sur les HGT. Ces descripteurs, comme la taille des génomes, le contenu en G+C, l'utilisation du carbone, la température et le pH optimaux, etc. seraient soumis à une analyse de variance (ANOVA). Les résultats pourraient être comparés à ceux obtenus par Jain et al. (2003).

## 4.7 Hypothèses alternatives/complémentaires

Les gènes ne sont pas toujours transférés individuellement, mais sous forme d'opéron ou de plasmide par exemple. Ce paramètre pourrait affecter la probabilité de fixation du matériel génétique transféré. En effet, si plusieurs gènes partenaires sont transférés simultanément, ils seront plus efficaces (et plus facilement fixés) qu'un gène seul tout juste arrivé dans un nouvel environnement génomique (Lawrence et Roth (1996), mais voir Pal et Hurst (2004) et Price et al. (2005)). Cependant notre approche ne se préoccupe pas de cette question, et nous n'avons aucun moyen de discriminer entre les deux cas de figures.

En plus de la rareté d'un gène, un autre critère pourrait influencer la propension au HGT : il s'agit de la plasticité du gène pour acquérir une nouvelle fonction. Imaginons un gène A codant pour une enzyme qui arrive par HGT chez une nouvelle espèce possédant déjà le gène correspondant. Le gène A restera pendant un certain temps dans l'espèce hôte avant d'être probablement perdu. Si durant cette période, un petit nombre de mutations affectant A lui permettent de métaboliser un nouveau substrat, il apportera alors un avantage sélectif à l'hôte, et verra augmenter de façon radicale ses chances d'être fixé.

Enfin, une hypothèse contraire à la notre a été formulée par Nicolas Rodrigue : le fait qu'un gène soit répandu devrait augmenter son taux de HGT simplement par effet de nombre. De plus, en cas de transfert, un tel gène aurait plus de chances de trouver des partenaires génomiques adéquats qu'un gène rare.

## 4.8 Conclusion

À partir de l'hypothèse relativement simple et vérifiée qu'était la nôtre, nous avons découvert un grand nombre de paramètres sous-jacents expliquant certaines particularités dans nos résultats, comme par exemple les singletons, ou l'artéfact de monophylie pour les familles de taille 4. Cependant, comme nous l'avons vu dans ce chapitre, des questions encore plus nombreuses ont été soulevées, nécessitant chacune un nouveau protocole pour y répondre.

## BIBLIOGRAPHIE


SF Altschul, W Gish, W Miller, EW Myers et DJ Lipman. Basic local alignment search tool. *J Mol Biol.*, 215(3):403–10, 1990.

SF Altschul, TL Madden, AA Schaffer, J Zhang, Z Zhang, W Miller et DJ Lipman. Gapped blast and psi-blast : a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–402, 1997.

L Aravind, RL Tatusov, YI Wolf, DL Walker et Koonin. Evidence for a massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet.*, 14:442–44, 2001.

S Aris-Brosou. Determinants of adaptive evolution at the molecular level : the extended complexity hypothesis. *Mol Biol Evol.*, 22(2):200–9, 2005.

OT Avery, CM MacLeod et M McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J. Exp. Med.*, 79(2): 137–58, 1944.

E Bapteste, Y Boucher, J Leigh et WF Doolittle. Phylogenetic reconstruction and lateral gene transfer. *Trends Microbiol.*, 12(9):406–11, 2004.

RG Beiko, TJ Harlow et MA Ragan. Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci USA*, 102(40):14332–7, 2005.

OG Berg et CG Kurland. Evolution of microbial genomes : sequence acquisition and loss. *Mol Biol Evol.*, 19(12):2265–76, 2002.

H Brinkmann, M van der Giezen, Y Zhou, G Poncelin de Raucourt et H Philippe. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst Biol.*, 54(5):743–57, 2005.

C Brochier, E Bapteste, D Moreira et Philippe H. Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet.*, 18(1):1–5, 2002.

J Castresana. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.*, 17(4):540–52, 2000.

E Chatton. *Titre et travaux scientifiques (1906-1937) de Edouard Chatton.* Sottano, Seton, France, 1938.

GD Clarke, RG Beiko, MA Ragan et RL Charlebois. Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized blastp scores. *J Bacteriol.*, 184(8):2072–80, 2002.

ML Coleman, MB Sullivan, AC Martiny, C Steglich, K Barry, EF Delong et SW Chisholm. Genomic islands and the ecology and evolution of prochlorococcus. *Science*, 311(5768):1768–70, 2006.

DQ Cortez, A Lazcano et A Becerra. Comparative analysis of methodologies for the detection of horizontally transferred genes. *In Silico Biol.*, 5(5-6):581–92, 2005.

C Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* John Murray, London, facsimile of the 6th (and last) edition of 1872, cambridge, ma, 1964 : harvard univ. press 1902 édition, 1859.

C Darwin. *The Variation of Animals and Plants Under Domestication.* John Murray, London, 1868.

C Darwin. *The Descent of Man, and Selection in Relation to Sex.* John Murray, London, 1871.

V Daubin, E Lerat et G Perriere. The source of laterally transferred genes in bacterial genomes. *Genome Biol.*, 4(9):R57, 2003a.

V Daubin, NA Moran et H Ochman. Phylogenetics and the cohesion of bacterial genomes. *Science*, 301(5634):829–832, 2003b.

V Daubin et G Perriere. G+c3 structuring along the genome : a common feature in prokaryotes. *Mol Biol Evol.*, 20(4):471–83, 2003.

B Dayrat. The roots of phylogeny how did haeckel build his trees? *Syst Biol.*, 52 (4):515–27, 2003.

MF DeFlaun et JH Paul. Detection of exogenous gene sequences in dissolved dna from aquatic environments. *Microb Ecol.*, 7:21–8, 1989.

T Dobzhansky. *Genetics and the Origin of Species.* 1937.

WF Doolittle. Phylogenetic classification and the universal tree. *Science*, 284 (5423):2124–9, 1999.

RC Edgar. Muscle : a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113, 2004a.

RC Edgar. Muscle : multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–7, 2004b.

J Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27:401–410, 1978.

J Felsenstein. *Inferring Phylogenies.* Sinauer Associates, Sunderland, MA, 2004.

J Felsenstein. Phylip (phylogeny inference package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle*, 2005.

RA Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Philosophical Transactions of the Royal Society of Edinburgh*, 52: 399Ű433, 1918.

WM Fitch et E Margoliash. Construction of phylogenetic trees. *Science*, 155(760): 279–84, 1967.

LS Frost, R Leplae, AO Summers et Toussaint A. Mobile genetic elements the agents of open source evolution. *Nat Rev Microbiol.*, 3(9):722–32, 2005.

DL Fulton, YY Li, MR Laird, BG Horsman, FM Roche et FS Brinkman. Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics*, 7: 270, 2006.

F Ge, LS Wang et J Kim. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. *PLoS Biol.*, 3(10):e310, 2005.

JP Gogarten, WF Doolittle et JG Lawrence. Prokaryotic evolution in light of gene transfer. *Mol Biol Evol.*, 19(12):2226–38, 2002.

JP Gogarten et JP Townsend. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol.*, 3(9):679–87, 2005.

A Gomez, T Galic, JF Mariet, I Matic, M Radman et MA Petit. Creating new genes by plasmid recombination in escherichia coli and bacillus subtilis. *Appl Environ Microbiol*, 71(11):7607–9, 2005.

FX Gomis-Ruth, M Sola, F de la Cruz et M Coll. Coupling factors in macromolecular type-iv secretion machineries. *Curr Pharm Des.*, 10(13):1551–65, 2004.

GS Gray et WM Fitch. Evolution of antibiotic resistance genes : the dna sequence of a kanamycin resistance gene from staphylococcus aureus. *Mol Biol Evol.*, 1 (1):57–66, 1983.

F Griffith. The significance of pneumococcal types. *J. Hyg.*, 27:113–159, 1928.

J Hacker, L Bender, M Ott, J Wingender, B Lund, R Marre et W Goebel. Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal escherichia coli isolates. *Microb Pathog.*, 8(3): 213–25, 1990.

J Hacker et JB Kaper. Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol.*, 54:641–79, 2000.

J Hacker, S Knapp et W Goebel. Spontaneous deletions and flanking regions of the chromosomally inherited hemolysin determinant of an escherichia coli o6 strain. *J Bacteriol.*, 154(3):1145–52, 1983.

E Haeckel. *Generelle morphologie der Organismen.* George Reimer, Berlin, Germany, 1866.

W Hao et GB Golding. Patterns of bacterial gene movement. *Mol Biol Evol.*, 21 (7):1294–307, 2004.

W Hao et GB Golding. The fate of laterally transferred genes life in the fast lane to adaptation or death. *Genome Res.*, 16(5):636–43, 2006.

TJ Harlow, JP Gogarten et MA Ragan. A hybrid clustering approach to recognition of protein families in 114 microbial genomes. *BMC Bioinformatics*, 5:545, 2004.

WS Hayes et M Borodovsky. How to interpret an anonymous bacterial genome machine learning approach to gene identification. *Genome Res.*, 8(11):1154–71, 1998.

DM Hillis et JJ Bull. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst Biol.*, 42(2):182–192, 1993.

H Ikeda, K Shiraishi et Y Ogata. Illegitimate recombination mediated by double-strand break and end-joining in escherichia coli. *Adv Biophys.*, 38:3–20, 2004.

R Jain, MC Rivera et JA Lake. Horizontal gene transfer among genomes : the complexity hypothesis. *Proc Natl Acad Sci USA*, 96(7):3801–6, 1999.

R Jain, MC Rivera, JE Moore et JA Lake. Horizontal gene transfer accelerates genome innovation and evolution. *Mol Biol Evol.*, 20(10):1598–602, 2003.

O Jeffroy, H Brinkmann, F Delsuc et H Philippe. Phylogenomics : the beginning of incongruence ? *Trends Genet.*, 22(4):225–31, 2006.

G Jobb, A von Haeseler et K Strimmer. Treefinder, a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol.*, 4:18, 2004.

D.A. Jonas, I Elmadfa, KH Engel, KJ Heller, G Kozianowski, A Konig, D Muller, JF Narbonne, W Wackernagel et J Kleiner. Safety considerations of dna in food. *Ann Nutr Metab*, 45(6):235–54, 2001.

T Kaneko et S Tabata. Complete genome structure of the unicellular cyanobacterium synechcystis sp. pcc6803. *Plant Cell Physiol.*, 38:1171–76, 1997.

DM Karl et MD Bailiff. The measurement and distribution of dissolved nucleic acids in aquatic environments. *Limnol. Oceanogr.*, 34:543–58, 1989.

M Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47:713–9, 1962.

LB Koski et GB Golding. The closest blast hit is often not the nearest neighbor. *J Mol Evol.*, 52(6):540–2, 2001.

CG Kurland. What tangled web : barriers to rampant horizontal gene transfer. *BioEssays*, 27(7):741–7, 2005.

CG Kurland, B Canback et OG Berg. Horizontal gene transfer : a critical view. *Proc Natl Acad Sci USA*, 100(17):9658–62, 2003.

JA Lake et MC Rivera. Deriving the genomic tree of life in the presence of horizontal gene transfer. *Mol Biol Evol.*, 21(4):681–90, 2004.

JG Lawrence et H Ochman. Amelioration of bacterial genomes : rates of change and exchange. *J Mol Evol.*, 44(4):383–97, 1997.

JG Lawrence et H Ochman. Reconciling the many faces of lateral gene transfer. *Trends Microbiol.*, 10(1):1–4, 2002.

JG Lawrence et JR Roth. Selfish operons : horizontal transfer may drive the evolution of gene clusters. *Genetics*, 143(4):1843–60, 1996.

LEDA. *version 4.4.1.* Algorithmic Solutions Software GmbH, Schuetzenstrasse 3 - 5, 66123 Saarbruecken, Germany.

Y Liu, PM Harrison, V Kunin et M Gerstein. Comprehensive analysis of pseudogenes in prokaryotes : widespread gene decay and failure of putative horizontally transferred genes. *Genome Biol.*, 5(9):R64, 2004.

PJ Lockhart, MA Steel, MD Hendy et D Penny. Recovering evolutionary trees under a more realistic model of sequence. *Mol Biol Evol.*, 11(4):605–12, 1994.

P Lopez, D Casane et Philippe H. Heterotachy, an important process of protein evolution. *Mol Biol Evol.*, 19(1):1–7, 2002.

MG Lorenz et W Wackernagel. Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol Rev.*, 58(3):563–602, 1994.

PR Marri, W Hao et GB Golding. Gene gain and gene loss in streptococcus : is it driven by habitat ? *Mol Biol Evol.*, 23(12):2379–91, 2006.

E Mayr. *Systematics and the Origin of Species.* 1942.

G Mendel. Versuche über plflanzen-hybriden. *Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr 1865*, pages 3–47, 1866.

TH Morgan, A Sturtevant, C Bridges et HJ Muller. *The Mechanism of Mendelian Inheritance.* 1915.

M Moscoso et JP Claverys. Release of dna into the medium by competent streptococcus pneumoniae kinetics, mechanism and stability of the liberated dna. *Mol Microbiol.*, 54(3):783–94, 2004.

AR Mushegian et EV Koonin. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci USA*, 93(19): 10004–6, 1996.

A Muto et S Osawa. The guanine and cytosine content of genomic dna and bacterial evolution. *Proc Natl Acad Sci USA*, 84(1):166–9, 1987.

Y Nakamura, T Itoh, H Matsuda et T Gojobori. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet.*, 36(7):760–6, 2004.

M Nei. Phylogenetic analysis in molecular evolutionary genetics. *Annu Rev Genet.*, 30:371–403, 1996.

KE Nelson, RA Clayton, SR Gill, ML Gwinn, RJ Dodson et *et al.* Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature*, 399:323–29, 1999.

AS Novozhilov, GP Karev et EV Koonin. Mathematical modeling of evolution of horizontally transferred genes. *Mol Biol Evol.*, 22(8):1721–32, 2005.

H Ochman et JG Lawrence. *Phylogenetics and the amelioration of bacterial genomes. In F. C. Neidhardt et al. (eds.) Escherichia coli and Salmonella typhimurium : Molecular and Cellular Biology*, chaptitre 141. ASM Publications, Washington, 2e édition, 1996.

H Ochman, E Lerat et V Daubin. Examining bacterial species under the specter of gene transfer and exchange. *Proc Natl Acad Sci USA*, 102(Suppl 1):6595–9, 2005.

H Ochman et AC Wilson. *Evolutionary history of enteric bacteria. In F. C. Neidhardt, ed. Escherichia coli and Salmonella typhimurium : Molecular and Cellular Biology.* ASM Press, Washington, D.C., 1987.

AV Ogram, GS Sayler et T Barkay. The extraction and purification of microbial dna from sediment. *J. Microbiol. Methods*, 7:57–66, 1987.

M Onda, J Yamaguchi, K Hanada, Y Asami et Ikeda H. Role of dna ligase in the illegitimate recombination that generates lambdabio-transducing phages in escherichia coli. *Genetics*, 158(1):29–39, 2001.

E Paget et Simonet. On the track of natural transformation in soil. *FEMS Microbiol Ecol.*, 15:109–17, 1994.

C Pal et LD Hurst. Evidence against the selfish operon theory. *Trends Genet.*, 20 (6):232–4, 2004.

R Palmen et KJ Hellingwerf. Uptake and processing of dna by acinetobacter calcoaceticus – a review. *Gene*, 192(1):179–90, 1997.

H Philippe, F Delsuc, H Brinkmann et N Lartillot. Phylogenomics. *Annu Rev Ecol Syst.*, 36(1):541–62, 2005.

H Philippe et CJ Douady. Horizontal gene transfer and phylogenetics. *Curr Opin Microbiol.*, 6(5):498–505, 2003.

CP Ponting, L Aravind, J Schultz, P Bork et EV Koonin. Eukaryotic signalling domain homologues in archaea and bacteria, ancient ancestry and horizontal gene transfer. *J Mol Biol.*, 289:729–45, 1999.

MN Price, KH Huang, AP Arkin et EJ Alm. Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res.*, 15(6):809–19, 2005.

MA Ragan. On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol Lett.*, 201(2):187–91, 2001.

MA Ragan et RL Charlebois. Distributional profiles of homologous open reading frames among bacterial phyla : implications for vertical and lateral transmission. *Int J Syst Evol Microbiol.*, 52(Pt 3):777–87, 2002.

MA Ragan, TJ Harlow et RG Beiko. Do different surrogate methods detect lateral genetic transfer events of different relative ages. *Trends Microbiol.*, 14(1):4–8, 2006.

MC Rivera, R Jain, JE Moore et JA Lake. Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci USA*, 95(11):6239–44, 1998.

B Roure, N Rodriguez-Ezpeleta et H Philippe. Scafos : a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evol Biol.*, 7(Suppl 1):S2, 2007.

MJ Sanderson et HB Shaffer. Troubleshooting molecular phylogentic analyses. *Annu Rev Ecol Evol Syst.*, 33:49–72, 2003.

J Sapp. The prokaryote-eukaryote dichotomy : Meanings and mythology. *Microbiol Mol Biol Rev.*, 69(2):292–305, 2005.

HA Schmidt, K Strimmer, M Vingron et A von Haeseler. Tree-puzzle : maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18(3):502–4, 2002.

PM Sharp et WH Li. The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, 15(3):1281–95, 1987.

MW Smith, DF Feng et RF Doolittle. Evolution by acquisition : the case for horizontal gene transfers. *Trends Biochem Sci.*, 17(12):489–93, 1992.

B Snel, P Bork et MA Huynen. Genomes in flux : the evolution of archaeal and proteobacterial gene content. *Genome Res.*, 12(1):17–25, 2002.

RY Stanier et CB van Niel. The concept of a bacterium. *Arch Mikrobiol.*, 42:17–35, 1962.

M Syvanen. Horizontal gene transfer : evidence and possible consequences. *Annu Rev Genet.*, 28:237–61, 1994.

RL Tatusov, EV Koonin et DJ Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–7, 1997.

CM Thomas et KM Nielsen. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol.*, 3(9):711–21, 2005.

JD Thompson, DG Higgins et TJ Gibson. Clustal w : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22(22), 1994.

A van Rijk et H Bloemendal. Molecular mechanisms of exon shuffling : illegitimate recombination. *Genetica*, 118(2-3):245–9, 2003.

CR Vossbrinck et CR Woese. Eukaryotic ribosomes that lack a 5.8s rna. *Nature*, 320(6059):287–8, 1986.

DP Wall, HB Fraser et AE Hirsh. Detecting putative orthologs. *Bioinformatics*, 19(13):1710–1, 2003.

RA Welch, V Burland, G 3rd Plunkett, P Redford, P Roesch, D Rasko, EL Buckles, SR Liou, A Boutin, J Hackett, D Stroud, GF Mayhew, DJ Rose, S Zhou, DC Schwartz, NT Perna, HL Mobley, MS Donnenberg et FR Blattner. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic escherichia coli. *Proc Natl Acad Sci USA*, 99(26):17020–4, 2002.

BM Wilkins et LS Frost. *Molecular Medical Microbiology.* Academic, London, 2001.

CR Woese. Bacterial evolution. *Microbiol Rev.*, 51(2):221–71, 1987.

CR Woese, L Achenbach, P Rouviere et L Mandelco. Archaeal phylogeny : reexamination of the phylogenetic position of archaeoglobus fulgidus in light of certain composition-induced artifacts. *Syst Appl Microbiol.*, 14(4):364–71, 1991.

CR Woese et GE Fox. Phylogenetic structure of the prokaryotic domain : the primary kingdoms. *Proc Natl Acad Sci USA*, 74(11):5088–90, 1977.

CR Woese, J Gibson et EG Fox. Do genealogical patterns in purple photosynthetic bacteria reflect interspecific gene transfer ? *Nature*, 283(5743):212–4, 1980.

CR Woese, O Kandler et Wheelis ML. Towards a natural system of organisms. *Proc Natl Acad Sci USA*, 87(12):4576–9, 1990.

O Zhaxybayeva, P Lapierre et JP Gogarten. Ancient gene duplications and the root(s) of the tree of life. *Protoplasma*, 227(1):53–64, 2006.

ND Zinder et J Lederberg. Genetic exchange in salmonella. *J Bacteriol.*, 64(5): 679–99, 1952.

E Zuckerkandl et L Pauling. Molecules as documents of evolutionary history. *J Theor Biol.*, 8(2):357–66, 1965.